

Dimensionality reduction for signal forecasting

Roman Isachenko

Skoltech advisor: Maxim Fedorov

MIPT advisor: Vadim Strijov

April 27, 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.

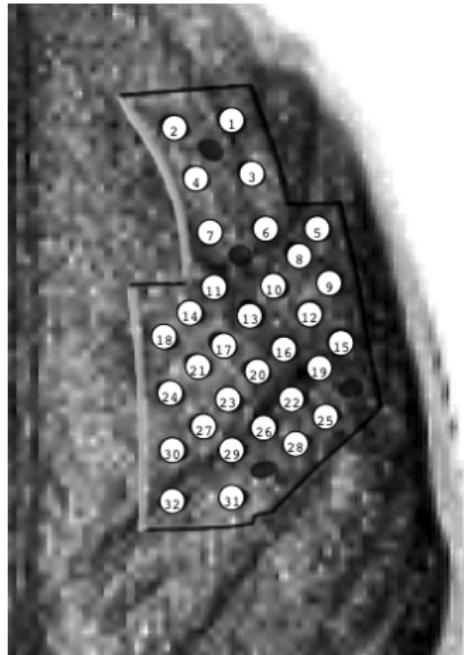
Solution

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.
- Li J. et al. Feature selection: A data perspective // *ACM Computing Surveys (CSUR)* 50(6), 2017.
- 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 $\mathbf{x} \in \mathbb{R}^n$.

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

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

$$\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_m]^T = [\chi_1, \dots, \chi_n]; \quad \mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_m]^T = [\nu_1, \dots, \nu_r].$$

Error function

$$S(\Theta | \mathbf{X}, \mathbf{Y}) = \left\| \mathbf{Y}_{m \times r} - \mathbf{X}_{m \times n} \cdot \Theta_{n \times r}^T \right\|_2^2 \rightarrow \min_{\Theta}.$$
$$\Theta = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}.$$

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 $\mathcal{A} = \{1, \dots, n\}$ 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

$$\mathcal{A} = \arg \max_{\mathcal{A}' \subseteq \{1, \dots, n\}} Q(\mathcal{A}' | \mathbf{X}, \mathbf{Y}).$$

Once the solution \mathcal{A} is known:

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

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

Quadratic Programming Feature Selection

$$\|\nu - \mathbf{X}\theta\|_2^2 \rightarrow \min_{\theta \in \mathbb{R}^n} .$$

Quadratic Programming Feature Selection

$$(1 - \alpha) \cdot \underbrace{\mathbf{a}^T \mathbf{Q} \mathbf{a}}_{\text{Sim}(\mathbf{X})} - \alpha \cdot \underbrace{\mathbf{b}^T \mathbf{a}}_{\text{Rel}(\mathbf{X}, \nu)} \rightarrow \min_{\substack{\mathbf{a} \geq \mathbf{0}_n \\ \mathbf{1}_n^T \mathbf{a} = 1}} .$$

- $\mathbf{a} \in \mathbb{R}^n$ — 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

$$\mathcal{A} = \arg \max_{\mathcal{A}' \subseteq \{1, \dots, n\}} Q(\mathcal{A}' | \mathbf{X}, \nu) \Leftrightarrow \arg \min_{\mathbf{a} \geq \mathbf{0}_n, \mathbf{1}_n^T \mathbf{a} = 1} [\mathbf{a}^T \mathbf{Q} \mathbf{a} - \alpha \cdot \mathbf{b}^T \mathbf{a}] .$$

Quadratic Programming Feature Selection

Quality Criteria

$$\mathcal{A} = \arg \max_{\mathcal{A}' \subseteq \{1, \dots, n\}} Q(\mathcal{A}' | \mathbf{X}, \nu) \Leftrightarrow \arg \min_{\mathbf{a} \geq \mathbf{0}_n, \mathbf{1}_n^T \mathbf{a} = 1} [\mathbf{a}^T \mathbf{Q} \mathbf{a} - \alpha \cdot \mathbf{b}^T \mathbf{a}].$$

Similarity measure

- Correlation

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

- Mutual information

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

$$\mathbf{Q} = \{|\text{corr}(\chi_i, \chi_j)|\}_{i,j=1}^n, \quad \mathbf{b} = \{|\text{corr}(\chi_i, \nu)|\}_{i=1}^n.$$

Statement

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

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

Multivariate QPFS

Relevance aggregation

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

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

Example:

$$\mathbf{X} = [\chi_1, \chi_2, \chi_3], \quad \mathbf{Y} = [\underbrace{\nu_1, \nu_1, \dots, \nu_1}_{r-1}, \nu_2],$$
$$\mathbf{Q} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0.8 \\ 0 & 0.8 & 1 \end{bmatrix}, \quad \mathbf{B} = \underbrace{\begin{bmatrix} 0.4 & \dots & 0.4 & 0 \\ 0.5 & \dots & 0.5 & 0.8 \\ 0.8 & \dots & 0.8 & 0.1 \end{bmatrix}}_{r-1}, \quad \mathbf{b} = \begin{bmatrix} (r-1) \cdot 0.4 + 0 \\ (r-1) \cdot 0.5 + 0.8 \\ (r-1) \cdot 0.8 + 0.1 \end{bmatrix}$$

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

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

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

Symmetric Importance

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

$$\alpha_1 \cdot \underbrace{\mathbf{a}_x^T \mathbf{Q}_x \mathbf{a}_x}_{\text{Sim}(\mathbf{X})} - \alpha_2 \cdot \underbrace{\mathbf{a}_x^T \mathbf{B} \mathbf{a}_y}_{\text{Rel}(\mathbf{X}, \mathbf{Y})} + \alpha_3 \cdot \underbrace{\mathbf{a}_y^T \mathbf{Q}_y \mathbf{a}_y}_{\text{Sim}(\mathbf{Y})} \rightarrow \min_{\substack{\mathbf{a}_x \geq \mathbf{0}_n, \mathbf{1}_n^T \mathbf{a}_x = 1 \\ \mathbf{a}_y \geq \mathbf{0}_r, \mathbf{1}_r^T \mathbf{a}_y = 1}}.$$

$$\mathbf{Q}_x = \{|\text{corr}(\chi_i, \chi_j)|\}_{i,j=1}^n, \quad \mathbf{Q}_y = \{|\text{corr}(\nu_i, \nu_j)|\}_{i,j=1}^r, \quad \mathbf{B} = \{|\text{corr}(\chi_i, \nu_j)|\}_{\substack{i=1, \dots, n \\ j=1, \dots, r}}.$$

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

Statement

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

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

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

Multivariate QPFS

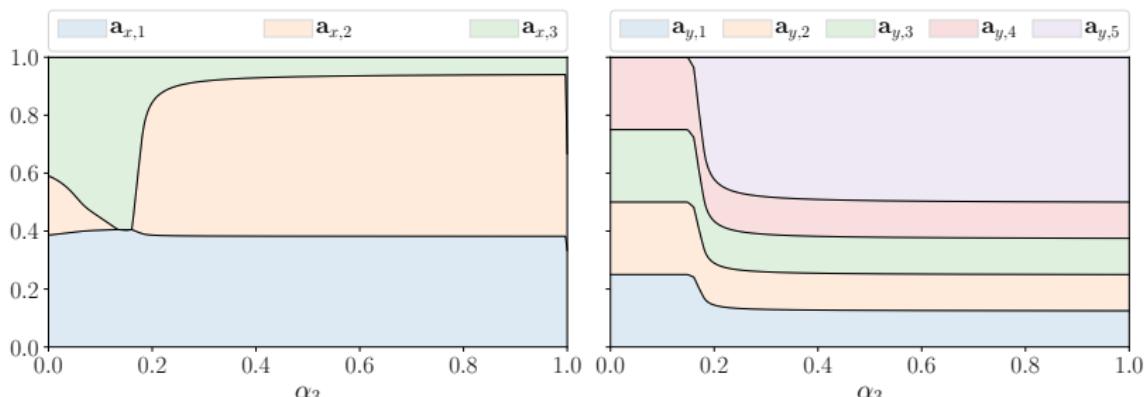
Example:

$$\mathbf{X} = [\chi_1, \chi_2, \chi_3], \quad \mathbf{Y} = [\underbrace{\nu_1, \nu_1, \dots, \nu_1}_{r-1}, \nu_2],$$

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

$$\mathbf{Q}_x = \mathbf{Q}; \quad \mathbf{Q}_y : \text{corr}(\nu_1, \nu_2) = 0.2 \text{ all others entries} = 1.$$

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



Min-max / Max-min

Symmetric Importances penalizes targets that are correlated and are not sufficiently explained by the features.

$$\alpha_1 \cdot \underbrace{\mathbf{a}_x^T \mathbf{Q}_x \mathbf{a}_x}_{\text{Sim}(\mathbf{X})} - \alpha_2 \cdot \underbrace{\mathbf{a}_x^T \mathbf{B} \mathbf{a}_y}_{\text{Rel}(\mathbf{X}, \mathbf{Y})} \rightarrow \min_{\substack{\mathbf{a}_x \geq \mathbf{0}_n, \\ \mathbf{1}_n^T \mathbf{a}_x = 1}} ; \quad \alpha_3 \cdot \underbrace{\mathbf{a}_y^T \mathbf{Q}_y \mathbf{a}_y}_{\text{Sim}(\mathbf{Y})} + \alpha_2 \cdot \underbrace{\mathbf{a}_x^T \mathbf{B} \mathbf{a}_y}_{\text{Rel}(\mathbf{X}, \mathbf{Y})} \rightarrow \min_{\substack{\mathbf{a}_y \geq \mathbf{0}_r, \\ \mathbf{1}_r^T \mathbf{a}_y = 1}} .$$

Min-max / Max-min

$$\min_{\substack{\mathbf{a}_x \geq \mathbf{0}_n \\ \mathbf{1}_n^T \mathbf{a}_x = 1}} \max_{\substack{\mathbf{a}_y \geq \mathbf{0}_r \\ \mathbf{1}_r^T \mathbf{a}_y = 1}} \left(\text{or} \max_{\substack{\mathbf{a}_y \geq \mathbf{0}_r \\ \mathbf{1}_r^T \mathbf{a}_y = 1}} \min_{\substack{\mathbf{a}_x \geq \mathbf{0}_n \\ \mathbf{1}_n^T \mathbf{a}_x = 1}} \right) \left[\alpha_1 \cdot \underbrace{\mathbf{a}_x^T \mathbf{Q}_x \mathbf{a}_x}_{\text{Sim}(\mathbf{X})} - \alpha_2 \cdot \underbrace{\mathbf{a}_x^T \mathbf{B} \mathbf{a}_y}_{\text{Rel}(\mathbf{X}, \mathbf{Y})} - \alpha_3 \cdot \underbrace{\mathbf{a}_y^T \mathbf{Q}_y \mathbf{a}_y}_{\text{Sim}(\mathbf{Y})} \right] .$$

$$\mathcal{A} = \arg \max_{\mathcal{A}' \subseteq \{1, \dots, n\}} Q(\mathcal{A}' | \mathbf{X}, \mathbf{Y}) \Leftrightarrow \arg \min_{\mathbf{a}_x \geq \mathbf{0}_n, \mathbf{1}_n^T \mathbf{a}_x = 1} \left[\max_{\substack{\mathbf{a}_y \geq \mathbf{0}_r \\ \mathbf{1}_r^T \mathbf{a}_y = 1}} f(\mathbf{a}_x, \mathbf{a}_y) \right] .$$

Theorem

For positive definite matrices \mathbf{Q}_x and \mathbf{Q}_y the max-min and min-max problems have the same optimal value.

Min-max / Max-min

Lagrangian for fixed \mathbf{a}_x

$$L(\mathbf{a}_x, \mathbf{a}_y, \lambda, \mu) = \alpha_1 \cdot \mathbf{a}_x^T \mathbf{Q}_x \mathbf{a}_x - \alpha_2 \cdot \mathbf{a}_x^T \mathbf{B} \mathbf{a}_y - \alpha_3 \cdot \mathbf{a}_y^T \mathbf{Q}_y \mathbf{a}_y + \lambda \cdot (\mathbf{1}_r^T \mathbf{a}_y - 1) + \mu^T \mathbf{a}_y.$$

Theorem

Min-max problem is equivalent to the following quadratic problem with $n + r + 1$ variables

$$\min_{\substack{\mathbf{a}_x \geq \mathbf{0}_n, \mathbf{1}_n^T \mathbf{a}_x = 1 \\ \lambda, \mu \geq \mathbf{0}_r}} g(\mathbf{a}_y, \lambda, \mu), \quad \text{where}$$

$$\begin{aligned} g(\mathbf{a}_x, \lambda, \mu) &= \max_{\mathbf{a}_y \in \mathbb{R}^r} L(\mathbf{a}_x, \mathbf{a}_y, \lambda, \mu) = \mathbf{a}_x^T \left(-\frac{\alpha_2^2}{4\alpha_3} \mathbf{B} \mathbf{Q}_y^{-1} \mathbf{B}^T - \alpha_1 \cdot \mathbf{Q}_x \right) \mathbf{a}_x \\ &\quad - \frac{1}{4\alpha_3} \lambda^2 \cdot \mathbf{1}_r^T \mathbf{Q}_y^{-1} \mathbf{1}_r - \frac{1}{4\alpha_3} \mu^T \mathbf{Q}_y^{-1} \mu + \frac{\alpha_2}{2\alpha_3} \lambda \cdot \mathbf{1}_r^T \mathbf{Q}_y^{-1} \mathbf{B}^T \mathbf{a}_x \\ &\quad - \frac{1}{2\alpha_3} \lambda \cdot \mathbf{1}_r^T \mathbf{Q}_y^{-1} \mu + \frac{\alpha_2}{2\alpha_3} \mu^T \mathbf{Q}_y^{-1} \mathbf{B}^T \mathbf{a}_x + \lambda. \end{aligned}$$

The problem is not convex. If we shift the spectrum for the matrix of quadratic form, the optimality is lost.

Minimax Relevances

Drop the term $\text{Sim}(\mathbf{Y})$.

$$\min_{\substack{\mathbf{a}_x \geq \mathbf{0}_n \\ \mathbf{1}_n^T \mathbf{a}_x = 1}} \max_{\substack{\mathbf{a}_y \geq \mathbf{0}_r \\ \mathbf{1}_r^T \mathbf{a}_y = 1}} \left[(1 - \alpha) \cdot \mathbf{a}_x^T \mathbf{Q}_x \mathbf{a}_x - \alpha \cdot \mathbf{a}_x^T \mathbf{B} \mathbf{a}_y \right].$$

Dual function

$$g(\mathbf{a}_x, \lambda, \mu) = \begin{cases} (1 - \alpha) \cdot \mathbf{a}_x^T \mathbf{Q}_x \mathbf{a}_x - \lambda, & \alpha \cdot \mathbf{B}^T \mathbf{a}_x = \lambda \cdot \mathbf{1}_r + \mu; \\ +\infty, & \text{otherwise.} \end{cases}$$

Statement

For the case $r = 1$ the proposed strategies coincide with the original QPFS algorithm.

Feature categorization

1. non-relevant features

$$\{j : \text{corr}(\chi_j, \nu_k) = 0, \forall k \in \{1, \dots, r\}\};$$

2. non-**X**-correlated features, which are relevant to non-**Y**-correlated targets

$$\{j : (\text{VIF}(\chi_j) < 10) \text{ and } (\text{VIF}(\nu_k) < 10, \forall k \in \{1, \dots, r\} : \text{corr}(\chi_j, \nu_k) \neq 0)\};$$

3. non-**X**-correlated features, which are relevant to **Y**-correlated targets

$$\{j : (\text{VIF}(\chi_j) < 10) \text{ and } (\exists k \in \{1, \dots, r\} : \text{VIF}(\nu_k) > 10 \& \text{corr}(\chi_j, \nu_k) \neq 0)\};$$

4. **X**-correlated features, which are relevant to non-**Y**-correlated targets

$$\{j : (\text{VIF}(\chi_j) > 10) \text{ and } (\text{VIF}(\nu_k) < 10, \forall k \in \{1, \dots, r\} : \text{corr}(\chi_j, \nu_k) \neq 0)\};$$

5. **X**-correlated features, which are relevant to **Y**-correlated targets

$$\{j : (\text{VIF}(\chi_j) > 10) \text{ and } (\exists k \in \{1, \dots, r\} : \text{VIF}(\nu_k) > 10 \& \text{corr}(\chi_j, \nu_k) \neq 0)\}.$$

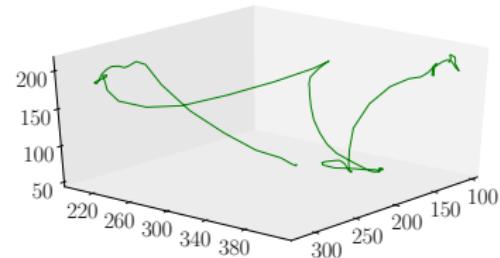
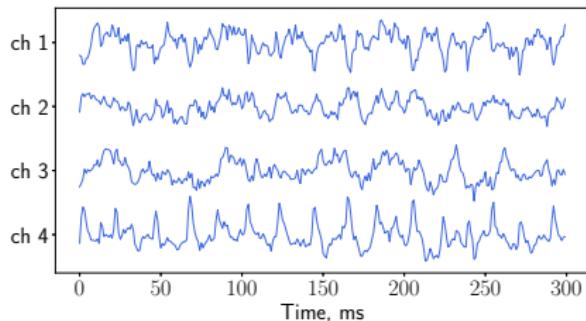
$$r = \text{corr}(\chi, \nu), \quad t = \frac{r\sqrt{m-2}}{1-r^2} \sim \text{St}(m-2).$$

$$\text{VIF}(\chi_j) = \frac{1}{1-R_j^2}, \quad \text{VIF}(\nu_k) = \frac{1}{1-R_k^2},$$

where R_j^2 (R_k^2) are coefficients of determination for the regression of χ_j (ν_k) on the other features(targets).

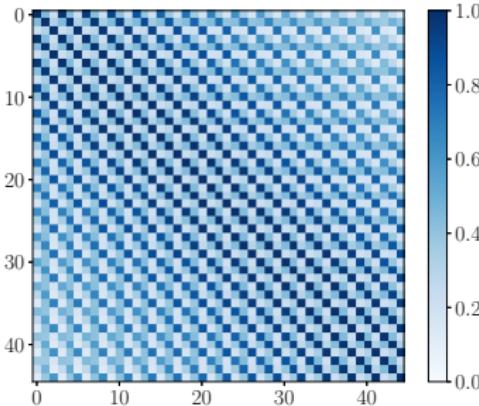
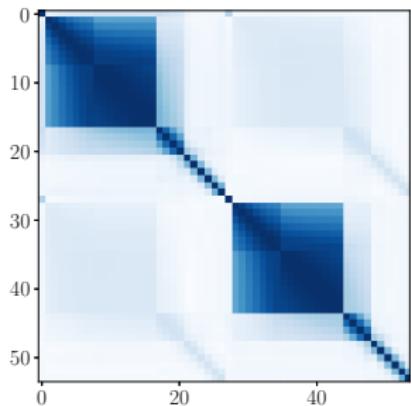
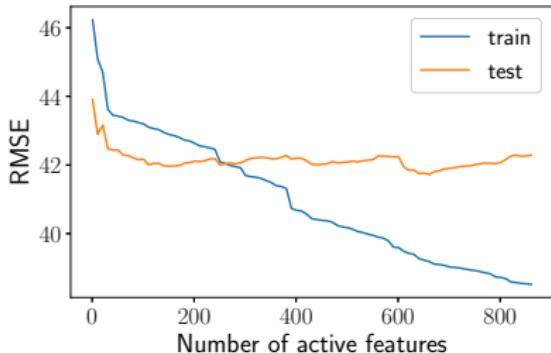
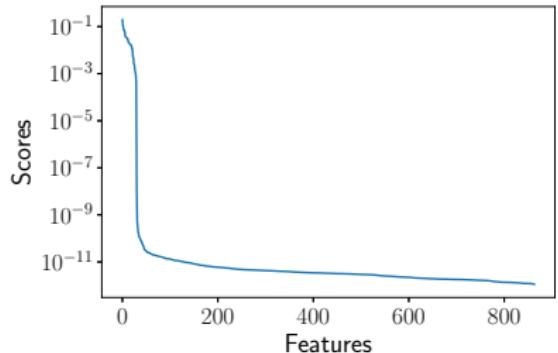
Computational experiment

Data example

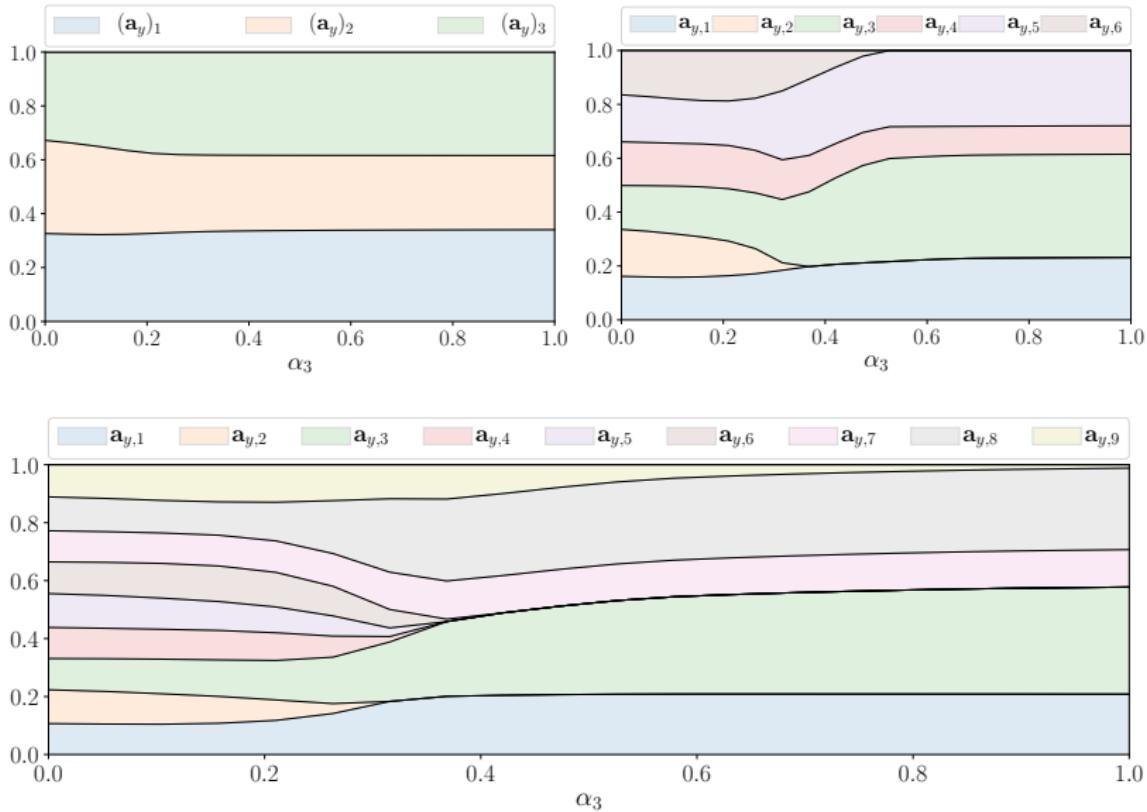


$$\mathbf{X} \in \mathbb{R}^{m \times (32 \cdot 27)}, \quad \mathbf{Y} = \begin{pmatrix} x_1 & y_1 & z_1 & \dots & x_k & y_k & z_k \\ x_2 & y_2 & z_2 & \dots & x_{k+1} & y_{k+1} & z_{k+1} \\ \dots & \dots & \dots & & \dots & \dots & \dots \\ x_m & y_m & z_m & \dots & x_{m+k} & y_{m+k} & z_{m+k} \end{pmatrix} \in \mathbb{R}^{m \times 3k}.$$

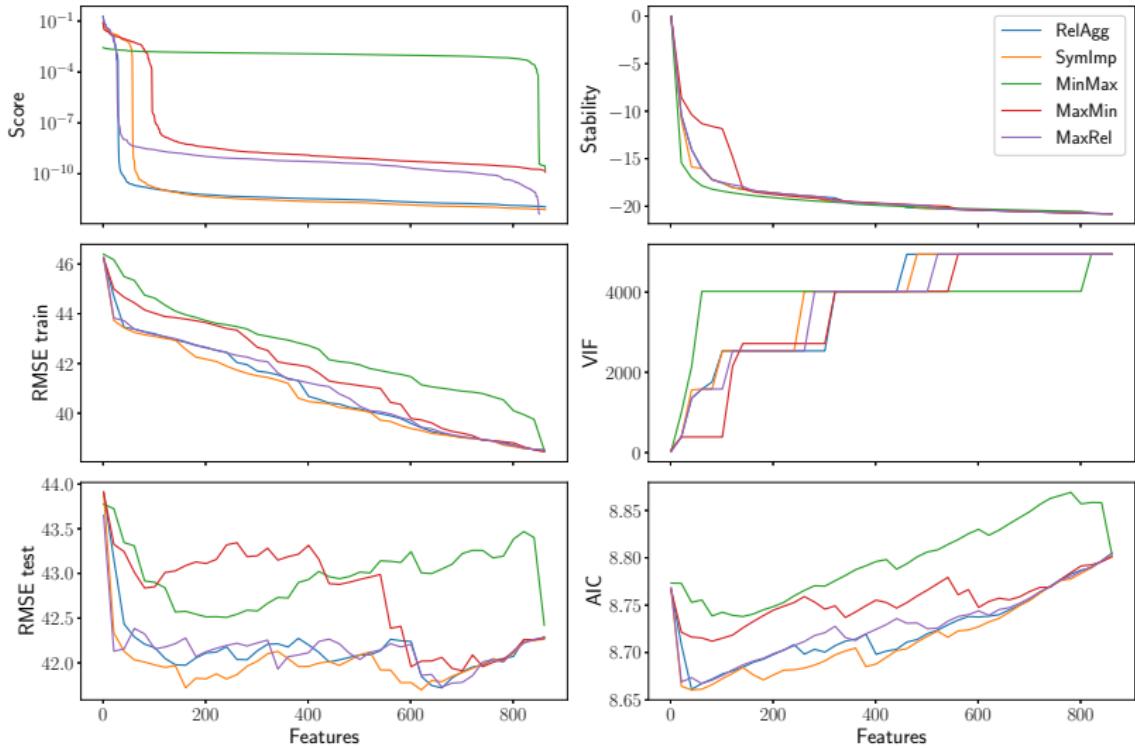
Experiment



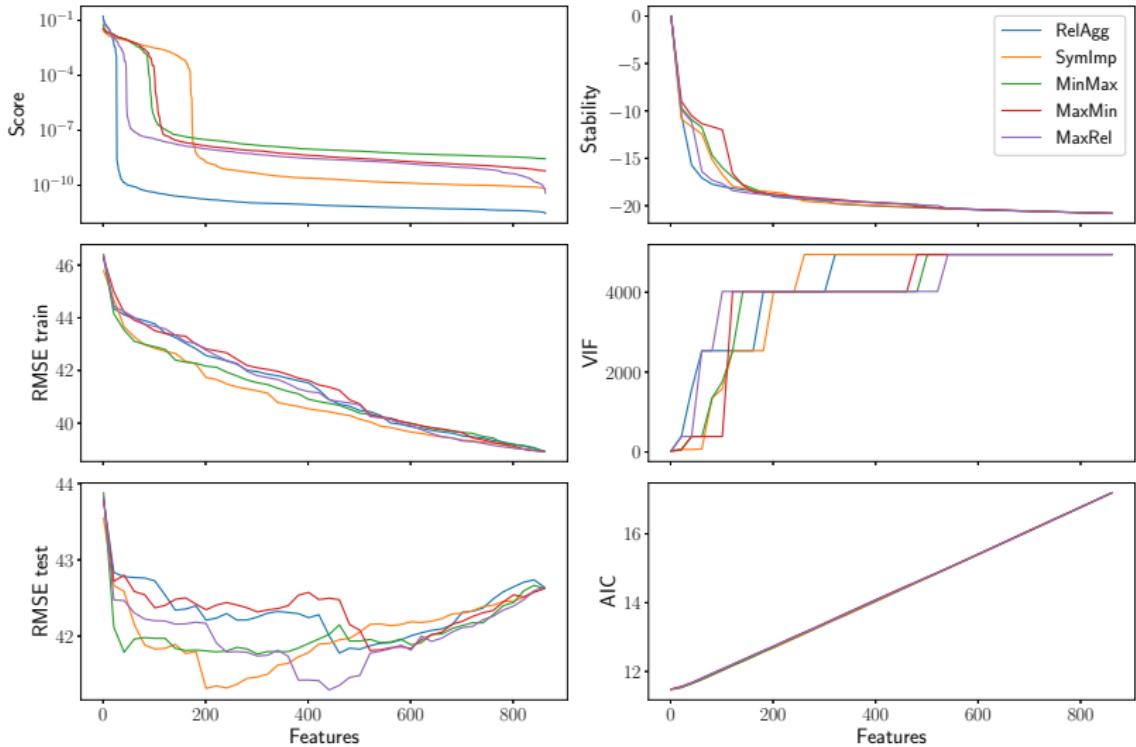
Experiment



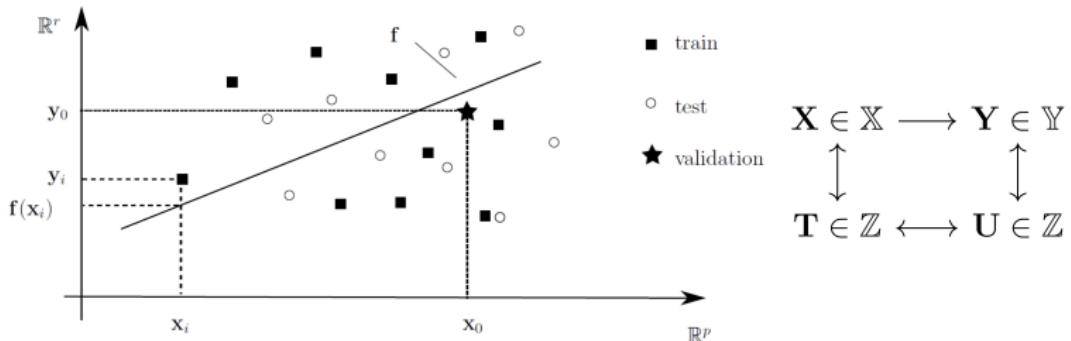
Autoregression step = 1



Autoregression step = 30



Problem Statement



Partial Least Squares (PLS)

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

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

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

- 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

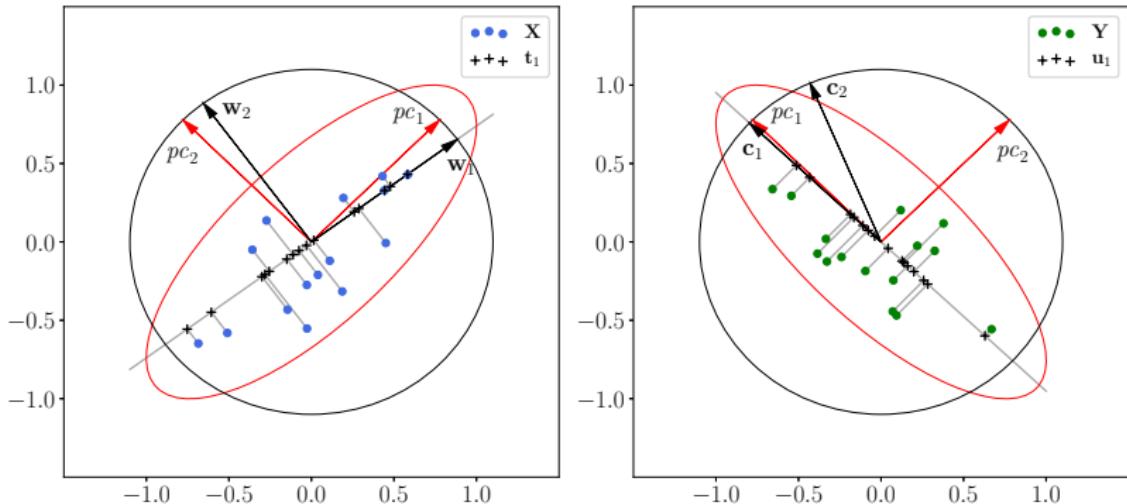


Figure: The result of the PLS algorithm for the case $n = r = l = 2$.

Computational experiment

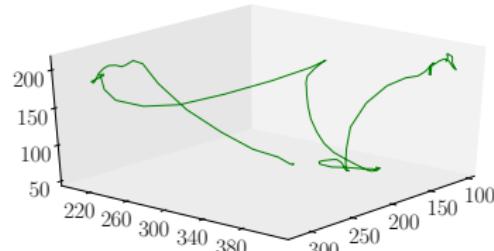
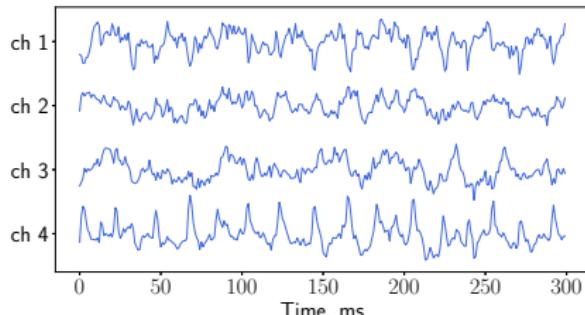
Datasets

- energy consumption
- electrocorticogram signals (ECoG)

Autoregressive approach

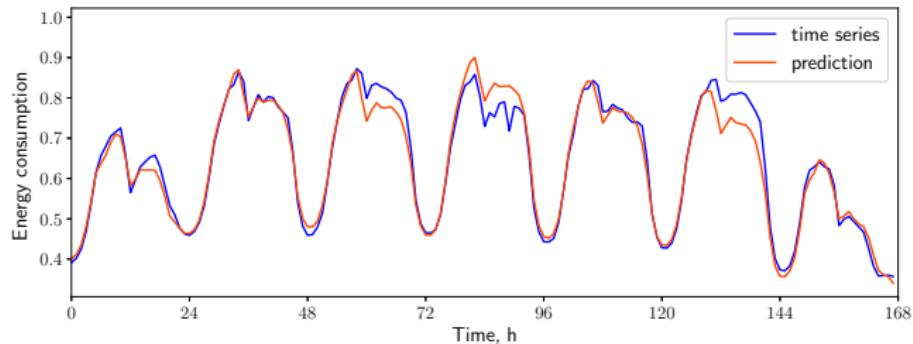
$$\mathbf{X} = \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ x_2 & x_3 & \dots & x_{n+1} \\ \dots & \dots & \dots & \dots \\ x_{T-n+1} & x_{T-n+2} & \dots & x_T \end{pmatrix}$$

ECoG data



Computational experiment

Energy consumption

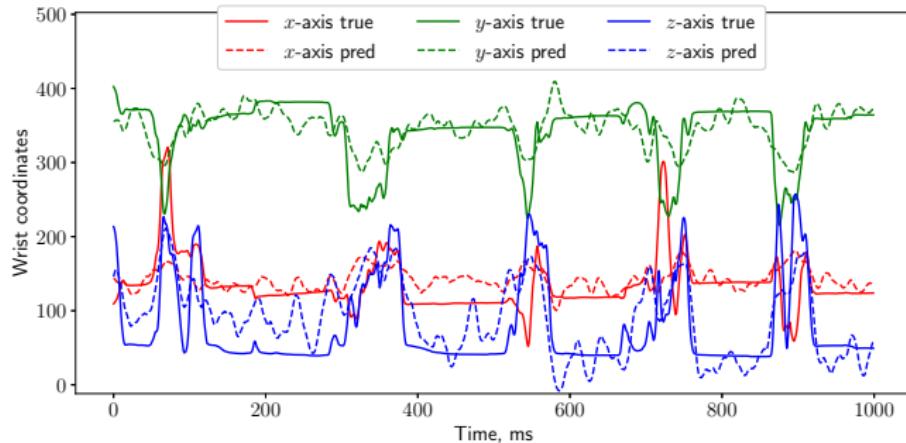


Results

- 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

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

- $\mathbf{z} \in \mathbb{R}^P$ — 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.

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

NN-QPMS

Model: feed-forward neural network

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

- weights from different layers do not interact;
- weight interactions → similarity between neurons;
- weight relevances → 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

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

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

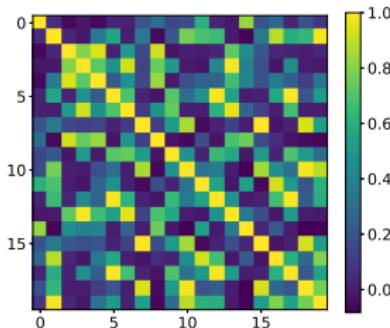


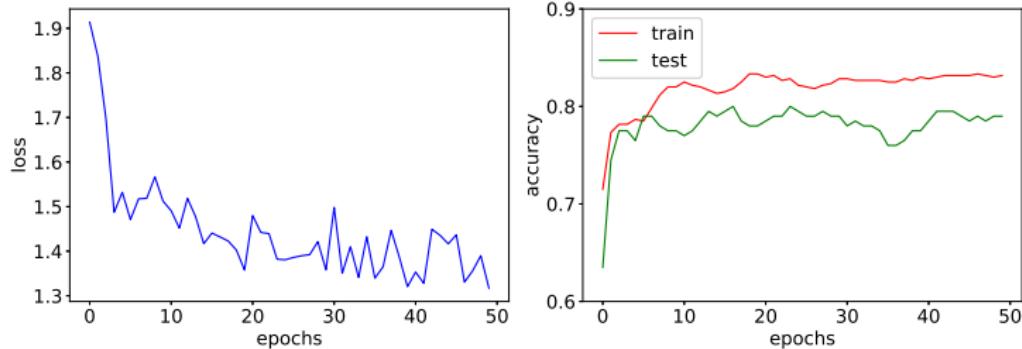
Figure: Correlation matrix

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

