Multivariate Quadratic Programming Feature Selection for ECoG signal decoding

R. V. Isachenko, V. V. Strijov

Abstract: The paper is devoted to the problem of signal decoding for Brain Computer Interface. The goal is to build a model which predicts the limb position by input brain signals. The challenge is redundancy in data description. The measurements are highly correlated. It leads to correlation in both input and target spaces. The feature selection are used to overcome the multicorrelation in feature representation. However, the majority of feature selection methods ignore the dependencies in the target space. Authors suggest the novel approach to the feature selection in the multivariate regression. The proposed methods extend the ideas of Quadratic Programming Feature Selection algorithm. The methods select non-correlated features that are relevant to the targets. The computational experiment shows the performance of the proposed algorithms in the ECoG data.

Keywords: multivariate regression, quadratic programming feature selection, multicorrelation

1 Introduction

The paper investigates the problem of decoding signal for Brain Computer Interface (BCI) [1]. The BCI aims to develop systems that help people with a severe motor control disability recover mobility. The minimally-invasive implant records cortical signals and the model decode them on real time to move the limbs of an exoskeleton [2, 3]. The subject placed inside the exoskeleton can drive it by imagining movements as if they were making the movement themself.

The challenge is redundancy of initial data description. The features are highly multicorrelated. The correlation comes from spatial nature of the data. The brain sensors are close to each other. It leads to the redundant measurements. In this case the final model is unstable. In addition, the redundant data description requires redundant computations which leads to real-time delay. To overcome this problem feature selection methods are used [4, 5].

One of the approach to the feature selection is to maximize feature relevances and minimize pairwise feature redundancy. This approach was proposed in the paper [6]. The

Quadratic Programming Feature Selection (QPFS) [7, 8] algorithm solves introduces two functions: Sim and Rel. The Sim measures the redundancy between features, the Rel contains relevances between each feature and the target vector. We want to minimize the function Sim and maximize the Rel simultaneously. QPFS offers the explicit way to construct the functions Sim and Rel. The method minimizes the following functional

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

The matrix $\mathbf{Q}\in\mathbb{R}^{n\times n}$ entries measure the pairwise similarities between features. The vector $\mathbf{b}\in\mathbb{R}^n$ expresses the similarities between each feature and the target matrix $\mathbf{b}$. The normalized vector $\mathbf{a}$ shows the importance of each feature. The functional (1) penalizes the dependent features by the function Sim and encourages features relevant to the target by the function Rel. The parameter α controls the trade-off between the functions Sim and the Rel. To measure similarity the authors use the absolute value of sample correlation coefficient or sample mutual information coefficient between pairs of features for the function Sim, and between features and the target vector for the function Rel.

We consider the multivariate problem, where the dependent variable is a vector. It refers to the prediction of limb position for not just one moment, but for some period of time. The subsequent hand position are correlated. It leads to correlations in the model output. In this situation feature selection algorithms do not take into account these dependencies. Hence, the selected feature subset is not optimal. We propose methods to take into account the dependencies in both input and output spaces. It allows to get the stable model with fewer variables.

The experiment was carried out in the ECoG data from the NeuroTycho project [9]. The proposed algorithms outperforms the original methods.

2 Problem statement

The goal is to forecast a dependent variable $\mathbf{y}\in\mathbb{R}^r$ with r targets from an independent input object $\mathbf{x}\in\mathbb{R}^n$ with n features. We assume there is a linear dependence

$$\mathbf{y}=\boldsymbol{\Theta}\mathbf{x}+\boldsymbol{\varepsilon} \tag{2}$$

between the objects $\mathbf{x}$ and the target variable $\mathbf{y}$, where $\boldsymbol{\Theta}\in\mathbb{R}^{r\times n}$ is the matrix of model parameters, $\boldsymbol{\varepsilon}\in\mathbb{R}^r$ is the residual vector. One has to find the matrix of the model parameters $\boldsymbol{\Theta}$ given a 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

$$\mathbf{X}=[\mathbf{x}_1,\ldots,\mathbf{x}_m]^{\mathsf{T}}=[\boldsymbol{\chi}_1,\ldots,\boldsymbol{\chi}_n];\quad \mathbf{Y}=[\mathbf{y}_1,\ldots,\mathbf{y}_m]^{\mathsf{T}}=[\boldsymbol{\nu}_1,\ldots,\boldsymbol{\nu}_r].$$

The columns $\boldsymbol{\chi}_j$ of the matrix $\mathbf{X}$ respond to object features.

The optimal parameters are determined by minimization of an error function. Define the quadratic loss function:

$$\mathcal{L}(\boldsymbol{\Theta}|\mathbf{X},\mathbf{Y}) = \left\| \underset{m\times r}{\mathbf{Y}} - \underset{m\times n}{\mathbf{X}} \cdot \underset{r\times n}{\boldsymbol{\Theta}^{\mathsf{T}}} \right\|_2^2 \to \min_{\boldsymbol{\Theta}}. \tag{3}$$

The solution of the problem (3) is given by

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

The linear dependent columns of the matrix $\mathbf{X}$ leads to an instable solution for the optimization problem (3). If there is a vector $\boldsymbol{\alpha} \neq \mathbf{0}_n$ such that $\mathbf{X}\boldsymbol{\alpha} = \mathbf{0}_m$, then adding the vector $\boldsymbol{\alpha}$ to any column of the matrix $\boldsymbol{\Theta}$ does not change the error function $S(\boldsymbol{\Theta}|\mathbf{X},\mathbf{Y})$. In this case the matrix $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ is not invertible. To avoid the strong linear dependence, feature selection and dimensionality reduction techniques are used.

3 Feature selection

The feature selection goal is to find the index set $\mathcal{A} = \{1,\ldots,n\}$ of the matrix $\mathbf{X}$ columns. To select the set $\mathcal{A}$ among all possible $2^n - 1$ subsets, introduce the feature selection error function

$$\mathcal{A} = \underset{\mathcal{A}'\subseteq\{1,\ldots,n\}}{\arg\min}\ S(\mathcal{A}'|\mathbf{X},\mathbf{Y}). \tag{4}$$

Once the solution $\mathcal{A}$ for the problem (4) is known, the problem (3) becomes

$$\mathcal{L}(\boldsymbol{\Theta}_{\mathcal{A}}|\mathbf{X}_{\mathcal{A}},\mathbf{Y}) = \left\|\mathbf{Y} - \mathbf{X}_{\mathcal{A}}\boldsymbol{\Theta}_{\mathcal{A}}^{\mathsf{T}}\right\|_2^2 \to \min_{\boldsymbol{\Theta}_{\mathcal{A}}}, \tag{5}$$

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

3.1 Quadratic Programming Feature Selection

The QPFS algorithm selects non-correlated features, which are relevant to the target vector $\boldsymbol{\nu}$ for the linear regression problem with $r = 1$

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

The QPFS functional 1 corresponds to the error function $S(\mathcal{A}|\mathbf{X},\boldsymbol{\nu})$

$$\mathcal{A} = \underset{\mathcal{A}'\subseteq\{1,\ldots,n\}}{\arg\min}\ S(\mathcal{A}'|\mathbf{X},\boldsymbol{\nu}) \Leftrightarrow \underset{\mathbf{a}\geq\mathbf{0}_n,\,\mathbf{1}_n^{\mathsf{T}}\mathbf{a}=1}{\arg\min}\ \left[\mathbf{a}^{\mathsf{T}}\mathbf{Q}\mathbf{a} - \alpha\cdot\mathbf{b}^{\mathsf{T}}\mathbf{a}\right].$$

The authors of the original QPFS paper suggested the way to select α and make $\mathrm{Sim}(\mathbf{X})$ and $\mathrm{Rel}(\mathbf{X},\boldsymbol{\nu})$ impact the same:

$$\alpha = \frac{\overline{\mathbf{Q}}}{\overline{\mathbf{Q}} + \overline{\mathbf{b}}},$$

where $\overline{\mathbf{Q}}$, $\overline{\mathbf{b}}$ are the mean values of $\mathbf{Q}$ and $\mathbf{b}$ respectively. Apply the thresholding for $\mathbf{a}$ to find the optimal feature subset:

$$\mathcal{A} = \big\{j \in \{1, \ldots, n\} : (\mathbf{a})_j > \tau\big\}.$$

We use the absolute value of sample correlation coefficient as similarity measure:

$$\mathbf{Q} = \big\{\big|\mathrm{corr}(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j)\big|\big\}_{i,j=1}^{n}, \quad \mathbf{b} = \{|\mathrm{corr}(\boldsymbol{\chi}_i, \boldsymbol{\nu})|\}_{i=1}^{n}. \tag{6}$$

The problem (1) is convex if the matrix $\mathbf{Q}$ is positive semidefinite. In general it is not always true. To satisfy this condition, the matrix $\mathbf{Q}$ spectrum is shifted and the matrix $\mathbf{Q}$ is replaced by $\mathbf{Q} - \lambda_{\min}\mathbf{I}$, where $\lambda_{\min}$ is a $\mathbf{Q}$ minimal eigenvalue.

3.2 Multivariate QPFS

Relevance aggregation (RelAgg). First approach to apply the QPFS algorithm to the multivariate case ($r > 1$) is to aggregate feature relevances through all r components. The term $\mathrm{Sim}(\mathbf{X})$ is still the same, and the matrix $\mathbf{Q}$ and the vector $\mathbf{b}$ are equal to

$$\mathbf{Q} = \big\{\big|\mathrm{corr}(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j)\big|\big\}_{i,j=1}^{n}, \quad \mathbf{b} = \left\{\sum_{k=1}^{r} |\mathrm{corr}(\boldsymbol{\chi}_i, \boldsymbol{\nu}_k)|\right\}_{i=1}^{n}.$$

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

$$\mathbf{X} = [\boldsymbol{\chi}_1, \boldsymbol{\chi}_2, \boldsymbol{\chi}_3], \quad \mathbf{Y} = [\underbrace{\boldsymbol{\nu}_1, \boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_1}_{r-1}, \boldsymbol{\nu}_2],$$

We have three features and r targets, where first $r-1$ target are the identical. The pairwise features similarities are given by the matrix $\mathbf{Q}$. The matrix $\mathbf{B}$ entries shows pairwise relevances features to the targets. The vector $\mathbf{b}$ is obtained by summation of the matrix $\mathbf{B}$ over columns.

$$\mathbf{Q} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0.8 \\ 0 & 0.8 & 1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0.4 & \ldots & 0.4 & 0 \\ 0.5 & \ldots & 0.5 & 0.8 \\ \underbrace{0.8 \quad \ldots \quad 0.8}_{r-1} & & & 0.1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} (r-1)\cdot 0.4 + 0 \\ (r-1)\cdot 0.5 + 0.8 \\ (r-1)\cdot 0.8 + 0.1. \end{bmatrix} \tag{7}$$

We would like to select only two features. For such configuration the best feature subset is $[\boldsymbol{\chi}_1, \boldsymbol{\chi}_2]$. The feature $\boldsymbol{\chi}_2$ predicts the second target $\boldsymbol{\nu}_2$ and the combination of features $\boldsymbol{\chi}_1, \boldsymbol{\chi}_2$ predicts the first component. The QPFS algorithm for $r = 2$ gives the solution $\mathbf{a} = [0.37, 0.61, 0.02]$. It coincides with our knowledge. However, if we add the collinear columns to the matrix $\mathbf{Y}$ and increase r to 5, the QPFS solution will be $\mathbf{a} = [0.40, 0.17, 0.43]$. Here we lost the relevant feature $\boldsymbol{\chi}_2$ and select the redundant feature $\boldsymbol{\chi}_3$.

Symmetric importances (SymImp). To take into account the dependencies in the columns of the matrix $\mathbf{Y}$ we extend the QPFS functional (1) to the multivariate case. We add the term $\mathrm{Sim}(\mathbf{Y})$ and extend the term $\mathrm{Rel}(\mathbf{X}, \mathbf{Y})$:

$$\alpha_1 \cdot \underbrace{\mathbf{a}_x^\mathsf{T} \mathbf{Q}_x \mathbf{a}_x}_{\mathrm{Sim}(\mathbf{X})} - \alpha_2 \cdot \underbrace{\mathbf{a}_x^\mathsf{T} \mathbf{B} \mathbf{a}_y}_{\mathrm{Rel}(\mathbf{X},\mathbf{Y})} + \alpha_3 \cdot \underbrace{\mathbf{a}_y^\mathsf{T} \mathbf{Q}_y \mathbf{a}_y}_{\mathrm{Sim}(\mathbf{Y})} \to \min_{\substack{\mathbf{a}_x \geq \mathbf{0}_n,\, \mathbf{1}_n^\mathsf{T} \mathbf{a}_x = 1 \\ \mathbf{a}_y \geq \mathbf{0}_r,\, \mathbf{1}_r^\mathsf{T} \mathbf{a}_y = 1}}. \tag{8}$$

Determine the entries of matrices $\mathbf{Q}_x \in \mathbb{R}^{n \times n}$, $\mathbf{Q}_y \in \mathbb{R}^{r \times r}$, $\mathbf{B} \in \mathbb{R}^{n \times r}$ in the following way

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

The vector $\mathbf{a}_x$ shows the feature importances, while $\mathbf{a}_y$ is a vector with the importance of each target. The targets which are correlated will be penalized by $\mathrm{Sim}(\mathbf{Y})$ and have the lower importances.

The coefficients α_1, α_2, and α_3 control the influence of each term to the functional (8) and satisfy the conditions:

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

Proposition 1. *The balance between the terms Sim*$(\mathbf{X})$*, Rel*$(\mathbf{X}, \mathbf{Y})$*, and Rel*$(\mathbf{X}, \mathbf{Y})$ *for the problem* (8) *is achieved by the following coefficients:*

$$\alpha_1 = \frac{\overline{\mathbf{Q}}_y \overline{\mathbf{B}}}{\overline{\mathbf{Q}}_y \overline{\mathbf{B}} + \overline{\mathbf{Q}}_x \overline{\mathbf{Q}}_y + \overline{\mathbf{Q}}_x \overline{\mathbf{B}}}; \quad \alpha_2 = \frac{\overline{\mathbf{Q}}_x \overline{\mathbf{Q}}_y}{\overline{\mathbf{Q}}_y \overline{\mathbf{B}} + \overline{\mathbf{Q}}_x \overline{\mathbf{Q}}_y + \overline{\mathbf{Q}}_x \overline{\mathbf{B}}}; \quad \alpha_3 = \frac{\overline{\mathbf{Q}}_x \overline{\mathbf{B}}}{\overline{\mathbf{Q}}_y \overline{\mathbf{B}} + \overline{\mathbf{Q}}_x \overline{\mathbf{Q}}_y + \overline{\mathbf{Q}}_x \overline{\mathbf{B}}},$$

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

Proof. The desired values of α_1, α_2, and α_3 are given by solution of the following equations

$$\begin{aligned} &\alpha_1 + \alpha_2 + \alpha_3 = 1; \\ &\alpha_1 \overline{\mathbf{Q}}_x = \alpha_2 \overline{\mathbf{B}} = \alpha_3 \overline{\mathbf{Q}}_y. \end{aligned}$$

Here, the mean values $\overline{\mathbf{Q}}_x$, $\overline{\mathbf{B}}$, $\overline{\mathbf{Q}}_y$ of the corresponding matrices $\mathbf{Q}_x$, $\mathbf{B}$, and $\mathbf{Q}_y$ are the mean values of the terms $\mathrm{Sim}(\mathbf{X})$, $\mathrm{Rel}(\mathbf{X}, \mathbf{Y})$, and $\mathrm{Rel}(\mathbf{X}, \mathbf{Y})$. □

To investigate the impact of the term $\mathrm{Sim}(\mathbf{Y})$ on the functional (8), we balance the terms $\mathrm{Sim}(\mathbf{X})$ and $\mathrm{Rel}(\mathbf{X}, \mathbf{Y})$ by fixing the proportion between α_1 and α_2:

$$\alpha_1 = \frac{(1 - \alpha_3)\overline{\mathbf{B}}}{\overline{\mathbf{Q}}_x + \overline{\mathbf{B}}}; \quad \alpha_2 = \frac{(1 - \alpha_3)\overline{\mathbf{Q}}_x}{\overline{\mathbf{Q}}_x + \overline{\mathbf{B}}}; \quad \alpha_3 \in [0, 1]. \tag{9}$$

We apply the proposed algorithm to the discussed example (7). The given matrix $\mathbf{Q}$ corresponds to the matrix $\mathbf{Q}_x$. We additionally define the matrix $\mathbf{Q}_y$ by setting $\mathrm{corr}(\boldsymbol{\nu}_1, \boldsymbol{\nu}_2) = 0.2$ and all others entries to one. Figure 1 shows the importances of features $\mathbf{a}_x$ and targets $\mathbf{a}_y$ with respect to α_3 coefficient. If α_3 is small, the impact of all targets are almost equal and the feature $\boldsymbol{\chi}_3$ dominates the feature $\boldsymbol{\chi}_2$. When α_3 becomes larger than 0.2, the importance $(\mathbf{a}_y)_5$ of the target ϕ_5 grows up along with the importance of the feature $\boldsymbol{\chi}_2$.

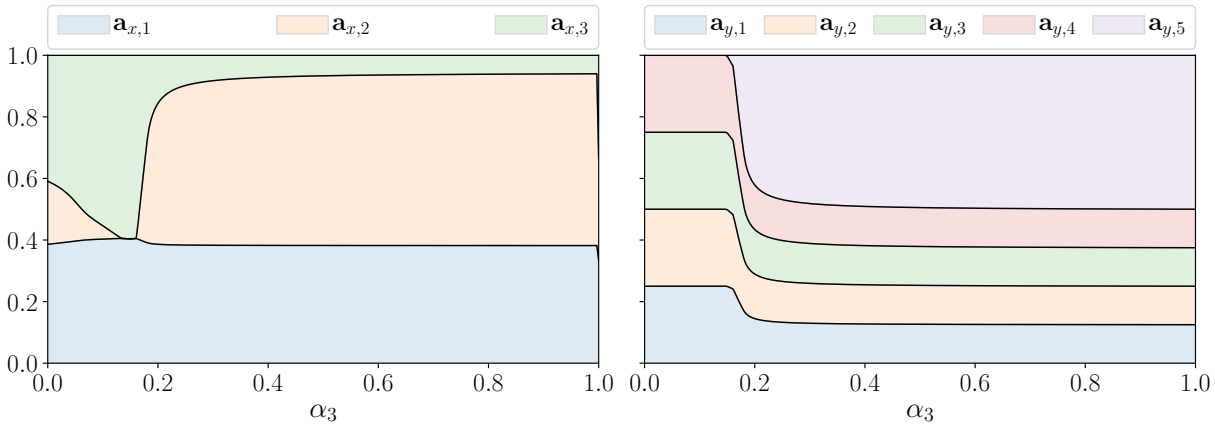


Figure 1: Feature importances $\mathbf{a}_x$ and $\mathbf{a}_y$ with respect to the α_3 coefficient

Minimax QPFS (MinMax and Maxmin). The functional (8) is symmetric with respect to $\mathbf{a}_x$ and $\mathbf{a}_y$. It penalizes features that are correlated and do not relevant to targets. At the same time it penalizes targets that are correlated and are not sufficiently explained by the features. It leads to small importances for targets which are difficult to predict by features and large importances for targets which are strongly correlated with features. It contradicts with the intuition. Our goal is to predict all targets, especially which are difficult to explain, by selected relevant and non-correlated features. We express this into two related problems.

$$\alpha_1 \cdot \underbrace{\mathbf{a}_x^\mathsf{T}\mathbf{Q}_x\mathbf{a}_x}_{\mathrm{Sim}(\mathbf{X})} - \alpha_2 \cdot \underbrace{\mathbf{a}_x^\mathsf{T}\mathbf{B}\mathbf{a}_y}_{\mathrm{Rel}(\mathbf{X},\mathbf{Y})} \to \min_{\substack{\mathbf{a}_x \geq \mathbf{0}_n, \\ \mathbf{1}_n^\mathsf{T}\mathbf{a}_x = 1}}; \tag{10}$$

$$\alpha_3 \cdot \underbrace{\mathbf{a}_y^\mathsf{T}\mathbf{Q}_y\mathbf{a}_y}_{\mathrm{Sim}(\mathbf{Y})} + \alpha_2 \cdot \underbrace{\mathbf{a}_x^\mathsf{T}\mathbf{B}\mathbf{a}_y}_{\mathrm{Rel}(\mathbf{X},\mathbf{Y})} \to \min_{\substack{\mathbf{a}_y \geq \mathbf{0}_r, \\ \mathbf{1}_r^\mathsf{T}\mathbf{a}_y = 1}}. \tag{11}$$

The difference in Rel part. In feature space the non-relevant components should have smaller scores. Meanwhile, the targets that are not relevant to the features should have larger scores. The problems (10) and (11) are merged into the joint min-max or max-min formulation

$$\min_{\substack{\mathbf{a}_x \geq \mathbf{0}_n \\ \mathbf{1}_n^\mathsf{T}\mathbf{a}_x = 1}} \max_{\substack{\mathbf{a}_y \geq \mathbf{0}_r \\ \mathbf{1}_r^\mathsf{T}\mathbf{a}_y = 1}} f(\mathbf{a}_x, \mathbf{a}_y), \quad \left(\text{or} \max_{\substack{\mathbf{a}_y \geq \mathbf{0}_r \\ \mathbf{1}_r^\mathsf{T}\mathbf{a}_y = 1}} \min_{\substack{\mathbf{a}_x \geq \mathbf{0}_n \\ \mathbf{1}_n^\mathsf{T}\mathbf{a}_x = 1}} f(\mathbf{a}_x, \mathbf{a}_y) \right), \tag{12}$$

where

$$f(\mathbf{a}_x, \mathbf{a}_y) = \alpha_1 \cdot \underbrace{\mathbf{a}_x^\mathsf{T}\mathbf{Q}_x\mathbf{a}_x}_{\mathrm{Sim}(\mathbf{X})} - \alpha_2 \cdot \underbrace{\mathbf{a}_x^\mathsf{T}\mathbf{B}\mathbf{a}_y}_{\mathrm{Rel}(\mathbf{X},\mathbf{Y})} - \alpha_3 \cdot \underbrace{\mathbf{a}_y^\mathsf{T}\mathbf{Q}_y\mathbf{a}_y}_{\mathrm{Sim}(\mathbf{Y})}.$$

Theorem 1. *For positive definite matrices* $\mathbf{Q}_x$ *and* $\mathbf{Q}_y$ *the max-min and min-max problems* (12) *have the same optimal value.*

Proof. Denote

$$\mathbb{C}^n = \{\mathbf{a} : \mathbf{a} \geq \mathbf{0}_n,\ \mathbf{1}_n^\mathsf{T}\mathbf{a} = 1\}, \quad \mathbb{C}^r = \{\mathbf{a} : \mathbf{a} \geq \mathbf{0}_r,\ \mathbf{1}_r^\mathsf{T}\mathbf{a} = 1\};$$

The sets $\mathbb{C}^n$ and $\mathbb{C}^r$ are compact and convex. The function $f : \mathbb{C}^n \times \mathbb{C}^r \to \mathbb{R}$ is a continuous function. If $\mathbf{Q}_x$ and $\mathbf{Q}_y$ are positive definite matrices, the function f is convex-concave, i.e. $f(\cdot, \mathbf{a}_y) : \mathbb{C}^n \to \mathbb{R}$ is convex for fixed $\mathbf{a}_y$, and $f(\mathbf{a}_x, \cdot) : \mathbb{C}^r \to \mathbb{R}$ is concave for fixed $\mathbf{a}_x$. In this case Neumann's minimax theorem states

$$\min_{\mathbf{a}_x \in \mathbb{C}^n} \max_{\mathbf{a}_y \in \mathbb{C}^r} f(\mathbf{a}_x, \mathbf{a}_y) = \max_{\mathbf{a}_y \in \mathbb{C}^r} \min_{\mathbf{a}_x \in \mathbb{C}^n} f(\mathbf{a}_x, \mathbf{a}_y).$$

□

To solve the min-max problem (12), fix some $\mathbf{a}_x \in \mathbb{C}^n$. For fixed vector $\mathbf{a}_x$ we solve the problem

$$\max_{\mathbf{a}_y \in \mathbb{C}_r} f(\mathbf{a}_x, \mathbf{a}_y) = \max_{\substack{\mathbf{a}_y \geq \mathbf{0}_r \\ \mathbf{1}_r^\mathsf{T}\mathbf{a}_y = 1}} \left[\alpha_1 \cdot \mathbf{a}_x^\mathsf{T}\mathbf{Q}_x\mathbf{a}_x - \alpha_2 \cdot \mathbf{a}_x^\mathsf{T}\mathbf{B}\mathbf{a}_y - \alpha_3 \cdot \mathbf{a}_y^\mathsf{T}\mathbf{Q}_y\mathbf{a}_y\right]. \tag{13}$$

The Lagrangian for this problem is

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

Here the Lagrange multipliers $\boldsymbol{\mu}$, corresponding to the inequality constraints $\mathbf{a}_y \geq \mathbf{0}_r$, are restricted to be nonnegative. The dual problem is

$$\min_{\lambda,\, \boldsymbol{\mu} \geq \mathbf{0}_r} g(\mathbf{a}_x, \lambda, \boldsymbol{\mu}) = \min_{\lambda,\, \boldsymbol{\mu} \geq \mathbf{0}_r} \left[\max_{\mathbf{a}_y \in \mathbb{R}^r} L(\mathbf{a}_x, \mathbf{a}_y, \lambda, \boldsymbol{\mu})\right]. \tag{14}$$

The strong duality holds for the problem (13). Therefore, the optimal value for (13) equals the optimal value for (14). It allows to solve the problem

$$\min_{\mathbf{a}_x \in \mathbb{C}^n,\, \lambda,\, \boldsymbol{\mu} \geq \mathbf{0}_r} g(\mathbf{a}_y, \lambda, \boldsymbol{\mu}) \tag{15}$$

instead of (12).

Setting the gradient of the Langrangian $\nabla_{\mathbf{a}_y} L(\mathbf{a}_x, \mathbf{a}_y, \lambda, \boldsymbol{\mu})$ to zero, we obtain an optimal value $\mathbf{a}_y$:

$$\mathbf{a}_y = \frac{1}{2\alpha_3}\mathbf{Q}_y^{-1}\left(-\alpha_2 \cdot \mathbf{B}^\mathsf{T}\mathbf{a}_x + \lambda \cdot \mathbf{1}_r + \boldsymbol{\mu}\right). \tag{16}$$

The dual function is equal to

$$\begin{aligned} g(\mathbf{a}_x, \lambda, \boldsymbol{\mu}) = \max_{\mathbf{a}_y \in \mathbb{R}^r} L(\mathbf{a}_x, \mathbf{a}_y, \lambda, \boldsymbol{\mu}) = \mathbf{a}_x^\mathsf{T}\left(-\frac{\alpha_2^2}{4\alpha_3}\mathbf{B}\mathbf{Q}_y^{-1}\mathbf{B}^\mathsf{T} - \alpha_1 \cdot \mathbf{Q}_x\right)\mathbf{a}_x \\ - \frac{1}{4\alpha_3}\lambda^2 \cdot \mathbf{1}_r^\mathsf{T}\mathbf{Q}_y^{-1}\mathbf{1}_r - \frac{1}{4\alpha_3}\boldsymbol{\mu}^\mathsf{T}\mathbf{Q}_y^{-1}\boldsymbol{\mu} + \frac{\alpha_2}{2\alpha_3}\lambda \cdot \mathbf{1}_r^\mathsf{T}\mathbf{Q}_y^{-1}\mathbf{B}^\mathsf{T}\mathbf{a}_x \\ - \frac{1}{2\alpha_3}\lambda \cdot \mathbf{1}_r^\mathsf{T}\mathbf{Q}_y^{-1}\boldsymbol{\mu} + \frac{\alpha_2}{2\alpha_3}\boldsymbol{\mu}^\mathsf{T}\mathbf{Q}_y^{-1}\mathbf{B}^\mathsf{T}\mathbf{a}_x + \lambda. \end{aligned} \tag{17}$$

It brings to the quadratic problem (15) with $n + r + 1$ variables.

Minimax Relevances (MaxRel). The problem (15) is not convex. If we shift the spectrum for the matrix of quadratic form (17), the optimality is lost. To overcome this problem, we drop the term $\mathrm{Sim}(\mathbf{Y})$.

$$\min_{\substack{\mathbf{a}_x \geq \mathbf{0}_n \\ \mathbf{1}_n^\mathsf{T}\mathbf{a}_x = 1}} \max_{\substack{\mathbf{a}_y \geq \mathbf{0}_r \\ \mathbf{1}_r^\mathsf{T}\mathbf{a}_y = 1}} \left[(1-\alpha)\cdot \mathbf{a}_x^\mathsf{T}\mathbf{Q}_x\mathbf{a}_x - \alpha \cdot \mathbf{a}_x^\mathsf{T}\mathbf{B}\mathbf{a}_y\right], \tag{18}$$

The Lagrangian for the problem (18) with the fixed vector $\mathbf{a}_x$ is

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

Setting the gradient of the Langrangian $\nabla_{\mathbf{a}_y} L(\mathbf{a}_x, \mathbf{a}_y, \lambda, \boldsymbol{\mu})$ to zero, we obtain:

$$\alpha \cdot \mathbf{B}^\mathsf{T}\mathbf{a}_x = \lambda \cdot \mathbf{1}_r + \boldsymbol{\mu}.$$

The dual function is equal to

$$g(\mathbf{a}_x, \lambda, \boldsymbol{\mu}) = \begin{cases} (1-\alpha)\cdot \mathbf{a}_x^\mathsf{T}\mathbf{Q}_x\mathbf{a}_x - \lambda, & \alpha \cdot \mathbf{B}^\mathsf{T}\mathbf{a}_x = \lambda \cdot \mathbf{1}_r + \boldsymbol{\mu}; \\ +\infty, & \text{otherwise.} \end{cases} \tag{19}$$

In this case the feature scores is the solution of (15).

Proposition 2. *For the case $r = 1$ the proposed functionals* (8) *and* (12) *coincide with the original QPFS algorithm* (1).

Proof. If r is equal to 1, then $\mathbf{Q}_y = q_y$ is a scalar, $\mathbf{a}_y = 1$, $\mathbf{B} = \mathbf{b}$. It reduces the problems (8) and (12) to

$$\alpha_1 \cdot \mathbf{a}_x^\mathsf{T}\mathbf{Q}_x\mathbf{a}_x - \alpha_2 \cdot \mathbf{a}_x^\mathsf{T}\mathbf{b} \to \min_{\mathbf{a}_x \geq \mathbf{0}_n,\, \mathbf{1}_n^\mathsf{T}\mathbf{a}_x = 1}.$$

Setting $\alpha = \frac{\alpha_2}{\alpha_1+\alpha_2}$ brings to the original QPFS problem (1). □

To summarize all proposed strategies for multivariate feature selection, Table 1 shows the core ideas and error functions for each method.

4 Metrics

To evaluate the selected subset $\mathcal{A}$ we introduce criteria that estimate the quality of feature selection procedure. We measure multicorrelation by mean value of miltiple correlation coefficient.

$$R^2 = \frac{1}{r}\mathrm{tr}\left(\mathbf{C}^\mathsf{T}\mathbf{R}^{-1}\mathbf{C}\right); \quad \text{where } \mathbf{C} = \{\mathrm{corr}(\boldsymbol{\chi}_i, \boldsymbol{\nu}_j)\}_{\substack{i=1,\ldots,n \\ j=1,\ldots,r}},\ \mathbf{R} = \{\mathrm{corr}(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j)\}_{i,j=1}^{n}.$$

This coefficient lies between 0 and 1. The bigger R^2 means the better feature subset we have.

Algorithm	Strategy	Error function $S(\mathcal{A}\|\mathbf{X},\mathbf{Y})$
RelAgg	$\min\big[\mathrm{Sim}(\mathbf{X}) - \mathrm{Rel}(\mathbf{X},\mathbf{Y})\big]$	$\min\limits_{\mathbf{a}_x}\big[(1-\alpha)\cdot\mathbf{a}_x^\mathsf{T}\mathbf{Q}_x\mathbf{a}_x - \alpha\cdot\mathbf{a}_x^\mathsf{T}\mathbf{B}\mathbf{1}_r\big]$
SymImp	$\min\big[\mathrm{Sim}(\mathbf{X}) - \mathrm{Rel}(\mathbf{X},\mathbf{Y}) + \mathrm{Sim}(\mathbf{Y})\big]$	$\min\limits_{\mathbf{a}_x,\,\mathbf{a}_y}\big[\alpha_1\cdot\mathbf{a}_x^\mathsf{T}\mathbf{Q}_x\mathbf{a}_x - \alpha_2\cdot\mathbf{a}_x^\mathsf{T}\mathbf{B}\mathbf{a}_y + \alpha_3\cdot\mathbf{a}_y^\mathsf{T}\mathbf{Q}_y\mathbf{a}_y\big]$
MinMax	$\min\big[\mathrm{Sim}(\mathbf{X}) - \mathrm{Rel}(\mathbf{X},\mathbf{Y})\big]$ $\max\big[\mathrm{Rel}(\mathbf{X},\mathbf{Y}) + \mathrm{Sim}(\mathbf{Y})\big]$	$\min\limits_{\mathbf{a}_x}\max\limits_{\mathbf{a}_y}\big[\alpha_1\cdot\mathbf{a}_x^\mathsf{T}\mathbf{Q}_x\mathbf{a}_x - \alpha_2\cdot\mathbf{a}_x^\mathsf{T}\mathbf{B}\mathbf{a}_y - \alpha_3\cdot\mathbf{a}_y^\mathsf{T}\mathbf{Q}_y\mathbf{a}_y\big]$
MaxMin	$\max\big[\mathrm{Rel}(\mathbf{X},\mathbf{Y}) + \mathrm{Sim}(\mathbf{Y})\big]$ $\min\big[\mathrm{Sim}(\mathbf{X}) - \mathrm{Rel}(\mathbf{X},\mathbf{Y})\big]$	$\max\limits_{\mathbf{a}_y}\min\limits_{\mathbf{a}_x}\big[\alpha_1\cdot\mathbf{a}_x^\mathsf{T}\mathbf{Q}_x\mathbf{a}_x - \alpha_2\cdot\mathbf{a}_x^\mathsf{T}\mathbf{B}\mathbf{a}_y - \alpha_3\cdot\mathbf{a}_y^\mathsf{T}\mathbf{Q}_y\mathbf{a}_y\big]$
MinRel	$\min\big[\mathrm{Sim}(\mathbf{X}) - \mathrm{Rel}(\mathbf{X},\mathbf{Y})\big]$ $\max\big[\mathrm{Rel}(\mathbf{X},\mathbf{Y})\big]$	$\min\limits_{\mathbf{a}_x}\max\limits_{\mathbf{a}_y}\big[(1-\alpha)\cdot\mathbf{a}_x^\mathsf{T}\mathbf{Q}_x\mathbf{a}_x - \alpha\cdot\mathbf{a}_x^\mathsf{T}\mathbf{B}\mathbf{a}_y\big]$

Table 1: Overview of proposed multivariate QPFS algorithms

The model stability is given by the logarithm of the ratio between minimal $\lambda_{\min}$ and maximum $\lambda_{\max}$ eigenvalue of the matrix $\mathbf{X}^\mathsf{T}\mathbf{X}$:

$$\mathrm{Stability} = \ln\frac{\lambda_{\min}}{\lambda_{\max}}.$$

The Root Mean Squared Error (RMSE) shows the quality of the model prediction. We estimate RMSE at the train and test data.

$$\mathrm{RMSE}(\mathbf{Y},\widehat{\mathbf{Y}}_{\mathcal{A}}) = \sqrt{\mathrm{MSE}(\mathbf{Y},\widehat{\mathbf{Y}}_{\mathcal{A}})} = \|\mathbf{Y}-\widehat{\mathbf{Y}}_{\mathcal{A}}\|_2, \quad \text{where} \quad \widehat{\mathbf{Y}}_{\mathcal{A}} = \mathbf{X}_{\mathcal{A}}\boldsymbol{\Theta}_{\mathcal{A}}^\mathsf{T}.$$

Akaike Information Criteria (AIC) is a trade-off between prediction quality and the size of selected subset $\mathcal{A}$:

$$\mathrm{AIC} = m\ln\left(\frac{\mathrm{MSE}(\mathbf{Y},\widehat{\mathbf{Y}}_{\mathcal{A}})}{m}\right) + 2|\mathcal{A}|.$$

5 Experiment

We carried out computational experiment with ECoG data from the NeuroTycho project. The input data consists of brain voltage signals recorded from 32 channels. The goal is to predict 3D hand position in the next moments given the input signal. The example of input signals and the 3D wrist coordinates are shown in Figure 2. The initial voltage signals are transformed to the spatial-temporal representation using wavelet transformation. The procedure of extracting feature representation from the raw data are described in details in [10,11]. Feature description in each time moment has dimension equals to 32 (channels) $\times$ 27 (frequencies) = 864. Each object is the representation of local history time segment

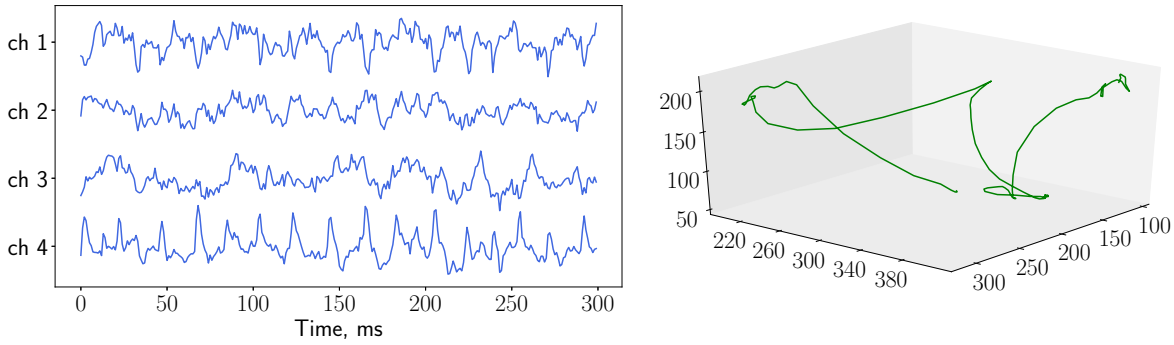


Figure 2: Brain signals and the corresponding hand position

with duration $\Delta t = 1s$. The time step between objects is $\delta t = 0.05s$. The final matrices are $\mathbf{X} \in \mathbb{R}^{18900 \times 864}$ and $\mathbf{Y} \in \mathbb{R}^{18900 \times 3k}$, where k is a number of timestamps that we predict. We split our data into train and test parts with the ratio 0.67.

Figures 3 and 4 show the result of the QPFS algorithm, where we use the Relevance Aggregation strategy and $k = 1$. QPFS scores $\mathbf{a}_x$ decrease sharply. Only about one hundred features have scores significantly greater that zero. The test error stops to decrease using more than this amount of features. It confirms that the initial data representation is highly redundant.

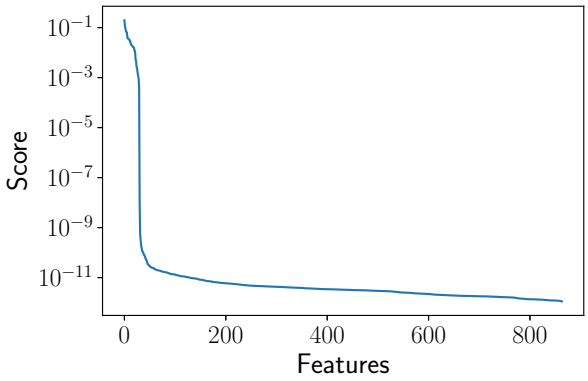


Figure 3: Sorted feature importances for the QPFS algorithm

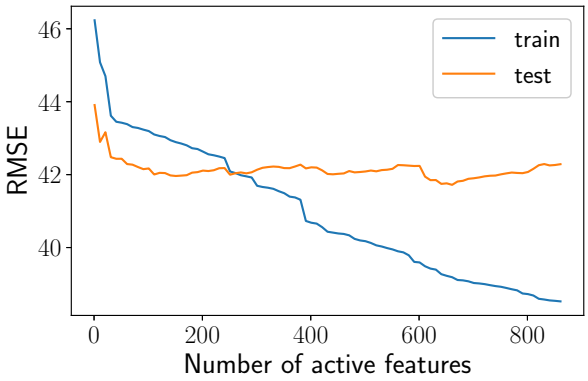


Figure 4: RMSE w.r.t. size of active set, features are ranked by QPFS algorithm

Figure 5 shows the dependencies in the matrices $\mathbf{X}$ and $\mathbf{Y}$. Frequencies in the matrix $\mathbf{X}$ are highly correlated. The correlations between axes are not significant in comparison with correlations between consequent moments.

We apply the QPFS algorithm with Relevance Aggregation strategy for different values of α_3 coefficient according to formulas (9). The dependence between target scores $\mathbf{a}_y$ with respect to α_3 for different values of k are shown in Figure 6. If we predict wrist coordinates only for one timestamp $k = 1$, targets scores are almost the same. It tells about the

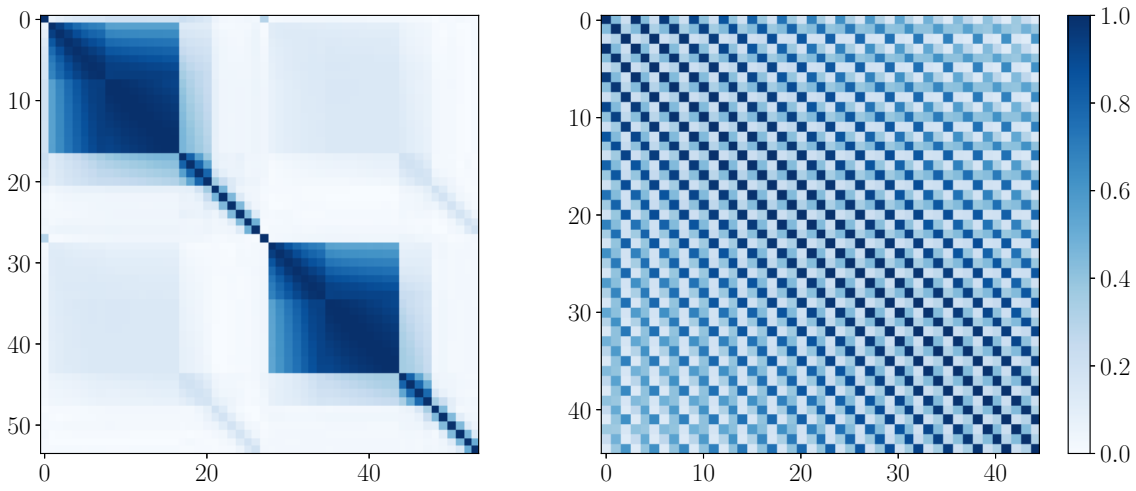


Figure 5: Correlation matrices for $\mathbf{X}$ and $\mathbf{Y}$

independence between x, y, and z coordinates. For $k = 2$ and $k = 3$ the scores of some targets becomes zero when α_3 increases.

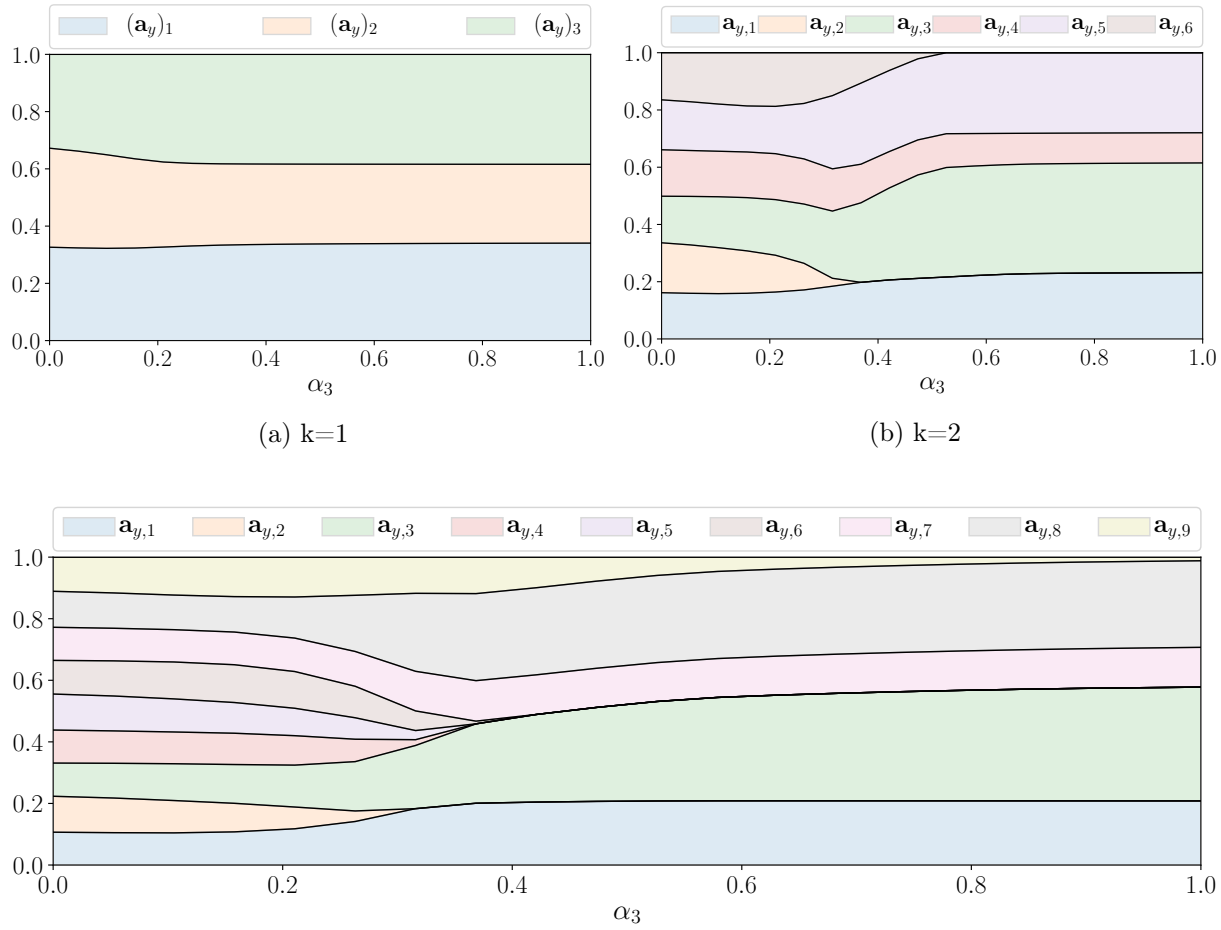
We compare the proposed strategies of multivariate QPFS that are given in Table 1 for the ECoG dataset. Firstly, we apply all methods to get feature scores. Then we fit linear model with increasing number of used features. For each method the features are sorted by the obtained scores. We show how the described metrics are changing with increasing the size of feature set. Figure 7 illustrates the results for $k = 1$. Here all metrics values for RelAgg, SymImp, and MaxRel are quite similar. However, MaxMin and MinMax algorithms show worse performance. This behavious is possibly due to unreasonable penalty on the matrix $\mathbf{Y}$.

The situation is changing for predicting several timestamps. In this case the correlation in $\mathbf{Y}$ matrix is crucial. Figure 8 shows algorithms performance for $k = 15$. The test error is minimal for SymImp, MinMax, and MaxRel strategies. The RelAgg strategy shows almost the worst error rate.

References

[1] Thomas Costecalde, Tetiana Aksenova, Napoleon Torres-Martinez, Andriy Eliseyev, Corinne Mestais, Cecile Moro, and Alim Louis Benabid. A long-term bci study with ecog recordings in freely moving rats. *Neuromodulation: Technology at the Neural Interface*, 21(2):149–159, 2018.

[2] Corinne S Mestais, Guillaume Charvet, Fabien Sauter-Starace, Michael Foerster, David Ratel, and Alim Louis Benabid. Wimagine: Wireless 64-channel ecog recording implant for long term clinical applications. *IEEE transactions on neural systems and rehabilitation engineering*, 23(1):10–21, 2015.



(a) k=1

(b) k=2

(c) k=3

Figure 6: Target importances $\mathbf{a}_y$ with respect to α_3 for QPFS with Relevance Aggregation

[3] Andrey Eliseyev, Corinne Mestais, Guillaume Charvet, Fabien Sauter, Neil Abroug, Nana Arizumi, Serpil Cokgungor, Thomas Costecalde, Michael Foerster, Louis Korczowski, et al. Clinatec® bci platform based on the ecog-recording implant wimagine® and the innovative signal-processing: preclinical results. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pages 1222–1225. IEEE, 2014.

[4] AM Katrutsa and VV Strijov. Stress test procedure for feature selection algorithms. *Chemometrics and Intelligent Laboratory Systems*, 142:172–183, 2015.

[5] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, 50(6):94, 2017.

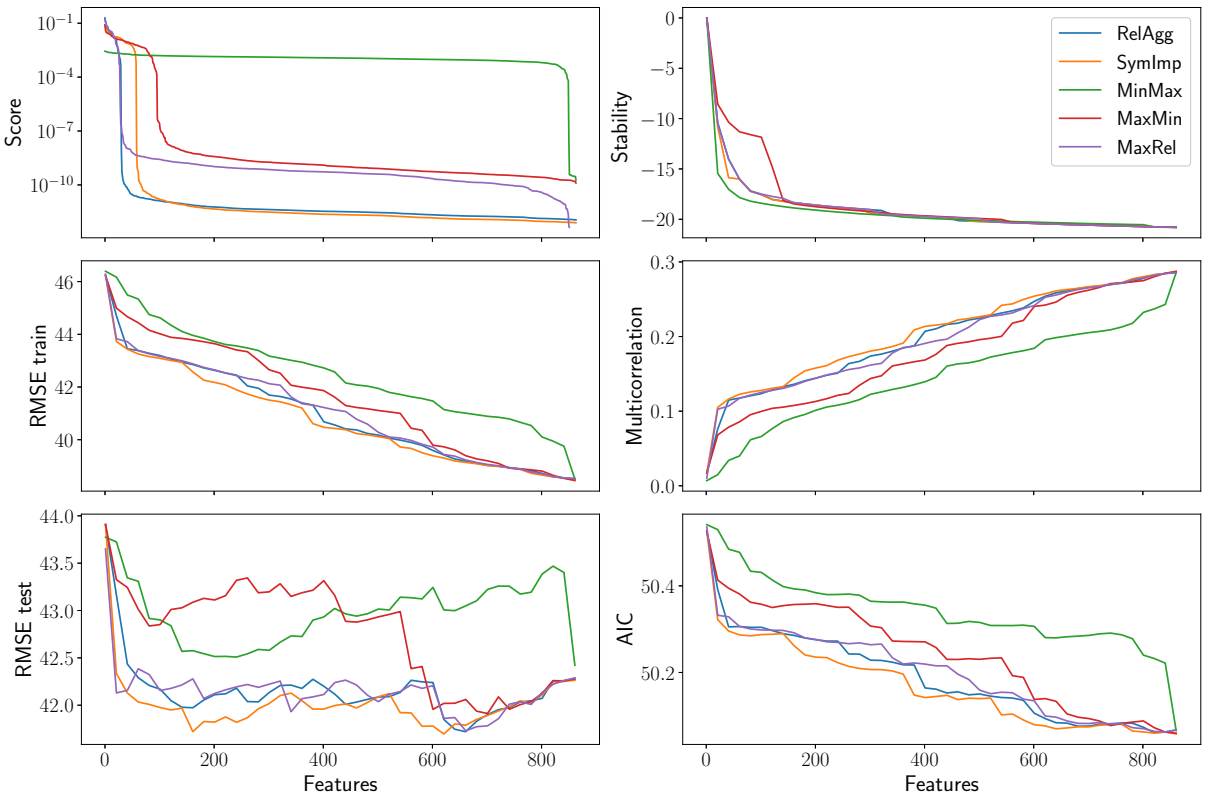


Figure 7: Metrics values for ECoG data with k = 1

[6] Chris Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3(02):185–205, 2005.

[7] Irene Rodriguez-Lujan, Ramon Huerta, Charles Elkan, and Carlos Santa Cruz. Quadratic programming feature selection. *Journal of Machine Learning Research*, 11(Apr):1491–1516, 2010.

[8] Alexandr Katrutsa and Vadim Strijov. Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria. *Expert Systems with Applications*, 76:1–11, 2017.

[9] Project tycho http://neurotycho.org/food-tracking-task.

[10] Zenas C Chao, Yasuo Nagasaka, and Naotaka Fujii. Long-term asynchronous decoding of arm motion using electrocorticographic signals in monkey. *Frontiers in neuroengineering*, 3:3, 2010.

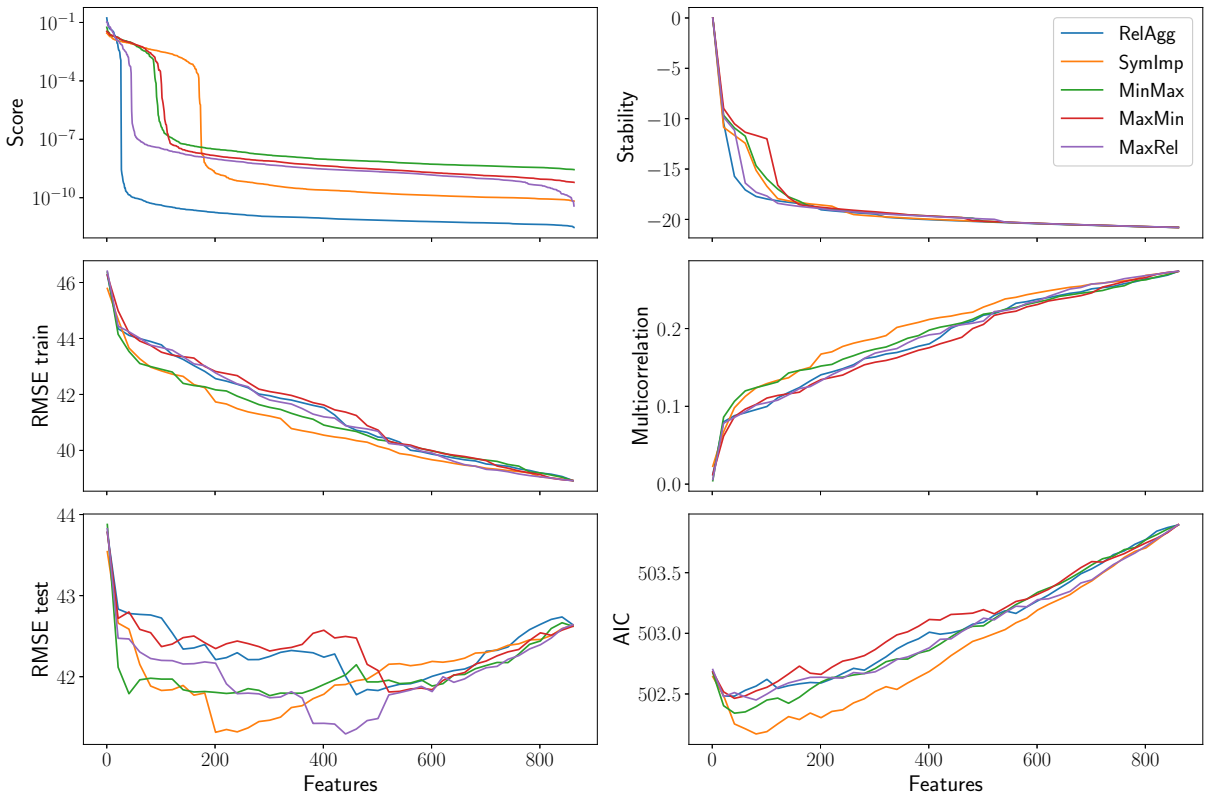


Figure 8: Metrics values for ECoG data with k = 15

[11] Andrey Eliseyev and Tetiana Aksenova. Penalized multi-way partial least squares for smooth trajectory decoding from electrocorticographic (ecog) recording. *PloS one*, 11(5):e0154878, 2016.