Multivariate Quadratic Programming Feature Selection for signal decoding

R. V. Isachenko, V. V. Strijov

Abstract: The paper is devoted to the problem of dimensionality reduction for signal decoding. The challenge of the investigation is redundancy in data description. High correlation in measurements leads to correlation in input space. The study considers multivariate case, the target variable is a vector. In this case the correlations occur in both input and target spaces. Dimensionality reduction and feature selection are used to build simple and stable model.

Partial least squares (PLS) regression is used as a base model for the dimensionality reduction. The model projects input and target data into the joint latent space and maximizes the covariances between the projections. To obrain the sparse model the feature selection is applied. The majority of feature selection methods ignore the dependencies in the target space. The study suggests a novel approach to feature selection in multivariate regression. The proposed approach extends the ideas of the quadratic programming feature selection (QPFS) algorithm. The QPFS algorithm selects non-correlated features, which are relevant to the targets. The proposed methods take into account the dependencies in the target space and select features which are informative to all targets jointly.

The computational experiment was carried out on the electrocorticogram (ECOG) and NIR spectrometry data. The proposed algorithms were compared by different criteria such as stability and their prediction performance. The algorithms give significantly better results compared to the baseline strategy. The linear regression with QPFS was compared with partial least squares (PLS) regression. The best result is obtained by combination of QPFS and PLS algorithms.

Keywords: multivariate regression, quadratic programming feature selection, signal decoding

1 Introduction

Initial data in the fields of chemometrics [1,2] and signal decoding [3,4] are high-dimensional and extremely redundant. The model which are built on such data are instable. In addition, the redundant data description requires excess computations which lead to real-time delay.

To overcome this problem dimensionality reduction [5,6] and feature selection [7,8] methods are used for high-dimensional data modelling.

Partial least squares (PLS) are widely used algorithm for dimensionality reduction [9–12]. PLS finds the optimal combinations of the initial features and use these combinations as the model features. The algorithm projects the features and the targets onto the joint latent space and maximizes the covariances between projected vectors. It allows to save information about initial input and target matrices and find their relations. The dimensionality of latent space is much less than the size of initial data description. It leads to a stable linear model built on the small number of features. The overview of advances in PLS regression is given in [13, 14]. For this model we obtain the linear model with small latent dimension. However, the final model use the whole range of the initial features and it does not allow to remove useless features.

Feature selection is a special case of dimensionality reduction when the latent representation is a subset of initial data description. Here the model are built on the subset of the features. One of the approach to feature selection is to maximize feature relevances and minimize pairwise feature redundancy. This approach was recently proposed and investigated in [15, 16]. Quadratic programmic feature selection (QPFS) [17] uses this approach to construct the optimization problem. It was shown in [18] that QPFS algorithm outperforms many existing feature selection methods for the univariate regression problem. The QPFS algorithm introduces two functions: Sim and Rel. Sim estimates the redundancy between features, Rel contains relevances between each feature and the target vector. QPFS minimizes the function Sim and maximizes the function Rel simultaneously. The algorithm solves the following optimization problem

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

Here columns of the matrix $\mathbf{X}$ are the features, and $\boldsymbol{\nu}$ is the target vector. 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 vector. The normalized vector $\mathbf{z}$ shows the importance of each feature. The function (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 Sim and the Rel. To measure similarity the authors use the absolute value of sample correlation coefficient between pairs of features for the function Sim, and between the features and the target vector for the function Rel.

The paper [19] proposes a multi-way version of the QPFS algorithm for tensor ECoG-based data. It was shown that QPFS is an appropriate feature selection method for signal decoding problem. We consider the multivariate problem, where the dependent variable is a vector. It leads to correlations in the model targets. In this situation feature selection algorithms do not take into account these dependencies. Hence, the selected feature subset is not optimal in terms of prediction. We propose methods which take into account the dependencies in both input and target spaces. It allows to get the stable sparse model. We refer to the original QPFS algortihm as our baseline for the computational experiment.

The main drawback of QPFS algorithm is computational cost. However, the original paper [17] suggests the way to solve the quadratic problem (1) efficiently. Additionally, in [20] the sequential minimal optimization framework is proposed for solving (1).

The experiments were carried out for ECoG dataset. We compared the proposed methods for multivariate feature selection with the baseline strategy and with PLS algorithm. The stability of the proposed methods were investigated by measuring how the feature selection solution changes with data bootstraping. The proposed algorithms outperform the baseline algorithm with the same number of features. The combination of the feature selection procedure and the PLS algorithm gives the best performance.

The main contributions of this paper are:

- addressing the dimensionality reduction problem for high-dimensional data;
- proposing new feature selection methods for multivariate regression
- comparing the proposed methods on real ECoG dataset, and showing that the proposed methods give the better feature subsets than the baseline method.

2 Multivariate regression

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 object $\mathbf{x}$ and the target variable $\mathbf{y}$, where $\boldsymbol{\Theta} \in \mathbb{R}^{r\times n}$ is a matrix of model parameters, $\boldsymbol{\varepsilon} \in \mathbb{R}^r$ is a 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

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

The columns $\boldsymbol{\chi}_j$ of $\mathbf{X}$ respond to the object features, the columns $\boldsymbol{\nu}_j$ of $\mathbf{Y}$ respond to the targets.

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

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

The solution of (3) is given by

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

The linear dependent columns of $\mathbf{X}$ 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 $\boldsymbol{\alpha}$ to any column of $\boldsymbol{\Theta}$ does not change the value of the loss function $\mathcal{L}(\boldsymbol{\Theta}|\mathbf{X}, \mathbf{Y})$. In this case the matrix $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ is close to singular and not invertible. To avoid the strong linear dependence, dimensionality reduction and feature selection are used.

3 Feature selection

The feature selection goal is to find the boolean vector $\mathbf{a} = \{0,1\}^n$, which components indicate whether the feature is selected. To obtain the optimal vector $\mathbf{a}$ among all possible $2^n - 1$ options, introduce the feature selection error function $S(\mathbf{a}|\mathbf{X},\mathbf{Y})$. We state the feature selection problem as follows

$$\mathbf{a} = \arg\min_{\mathbf{a}' \in \{0,1\}^n} S(\mathbf{a}'|\mathbf{X},\mathbf{Y}). \tag{4}$$

The goal of feature selection is to construct the appropriate function $S(\mathbf{a}|\mathbf{X},\mathbf{Y})$. The particular examples for the considered feature selection algorithms are given below and summarized in the Table 1.

The problem (4) are hard to solve due to discrete binary domain $\{0,1\}^n$. We relax the problem (4) to the continuous domain $[0,1]^n$. The relaxed feature selection problem is

$$\mathbf{z} = \arg\min_{\mathbf{z}' \in [0,1]^n} S(\mathbf{z}'|\mathbf{X},\mathbf{Y}). \tag{5}$$

Here the vector $\mathbf{z}$ entries are normalized feature importances. Firstly, solve the problem (5) to obtain the feature importances $\mathbf{z}$. Then the solution of (4) is recovered by thresholding:

$$\mathbf{a} = [a_j]_{j=1}^n, \quad a_j = \begin{cases} 1, & z_j > \tau; \\ 0, & \text{otherwise.} \end{cases}$$

Here the value τ is a hyperparameter which is defined manually or choosen by cross-validation.

Once the solution $\mathbf{a}$ of (4) is known, the problem (3) becomes

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

where the subscript $\mathbf{a}$ indicates the submatrix with the columns for which components of $\mathbf{a}$ equal 1.

3.1 Quadratic Programming Feature Selection

Our base algorithm for feature selection is quadratic programming feature selection algorithm. The paper [18] shows that QPFS outperforms many existing feature selection algorithms in different criteria. 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 authors of the original QPFS paper [17] suggested the way to select α for (1) and make $\text{Sim}(\mathbf{X})$ and $\text{Rel}(\mathbf{X},\boldsymbol{\nu})$ impacts 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. The QPFS parameters are defined as follows:

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

Here $\mathrm{corr}(\cdot, \cdot)$ is the absolute value of sample Pearson correlation coefficient

$$|\mathrm{corr}(\boldsymbol{\chi}, \boldsymbol{\nu})| = \left| \frac{\sum_{i=1}^{m} (\boldsymbol{\chi}_i - \overline{\boldsymbol{\chi}})(\boldsymbol{\nu}_i - \overline{\boldsymbol{\nu}})}{\sqrt{\sum_{i=1}^{m} (\boldsymbol{\chi}_i - \overline{\boldsymbol{\chi}})^2 \sum_{i=1}^{m} (\boldsymbol{\nu}_i - \overline{\boldsymbol{\nu}})^2}} \right|.$$

The other ways to define $\mathbf{Q}$ and $\mathbf{b}$ are considered in [18].

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 minimal eigenvalue of $\mathbf{Q}$.

3.2 Multivariate QPFS

We are aimed to propose the algorithms, which are suitable for feature selection in the multivariate case. If the target space is multidimensional it prone to redundancy and correlations between the targets. In this section the algorithms that take into account dependencies in both input and target spaces are proposed.

3.3 Relevance aggregation (RelAgg).

In [19] to apply QPFS algorithm for the multivariate case ($r > 1$) feature relevances are aggregated through all r components. The term $\mathrm{Sim}(\mathbf{X})$ is still the same, the matrix $\mathbf{Q}$ is defined by (6). The vector $\mathbf{b}$ is aggregated across all targets and is defined by

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

The drawback of this approach is its insensitivity to the dependencies in the columns of $\mathbf{Y}$. Observe the following example:

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

We have 3 features and r targets, where first $r - 1$ targets are identical. The pairwise features similarities are given by the matrix $\mathbf{Q}$. The matrix $\mathbf{B}$ entries show pairwise features relevances to the targets. The vector $\mathbf{b}$ is obtained by summation of the matrix $\mathbf{B}$ over the columns

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

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 $\mathbf{z} = [0.37, 0.61, 0.02]$. 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 $\mathbf{z} = [0.40, 0.17, 0.43]$. Here we lose the relevant feature $\boldsymbol{\chi}_2$ and select the redundant feature $\boldsymbol{\chi}_3$. The following subsections propose the extension of the QPFS algorithm which are overcome the challenge of this example.

3.4 Symmetric importances (SymImp).

To take into account the dependencies in the columns of the matrix $\mathbf{Y}$ we extend the QPFS function (1) to the multivariate case. We add the term $\mathrm{Sim}(\mathbf{Y})$ and modify the term $\mathrm{Rel}(\mathbf{X}, \mathbf{Y})$ as follows

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

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

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

The vector $\mathbf{z}_x$ shows the features importances, while $\mathbf{z}_y$ is a vector with the targets importances. The correlated targets 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 on the function (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 Sim*$(\mathbf{Y})$ *for the problem* (8) *is achieved by the following coefficients:*

$$\alpha_1 \propto \overline{\mathbf{Q}}_y \overline{\mathbf{B}}; \quad \alpha_2 \propto \overline{\mathbf{Q}}_x \overline{\mathbf{Q}}_y; \quad \alpha_3 \propto \overline{\mathbf{Q}}_x \overline{\mathbf{B}}. \tag{9}$$

Here $\overline{\mathbf{Q}}_x$*,* $\overline{\mathbf{B}}$*,* $\overline{\mathbf{Q}}_y$ *are 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 solving of the following equations

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

Here, the mean values $\overline{\mathbf{Q}}_x$, $\overline{\mathbf{B}}$, and $\overline{\mathbf{Q}}_y$ 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 $\mathrm{Sim}(\mathbf{Y})$. □

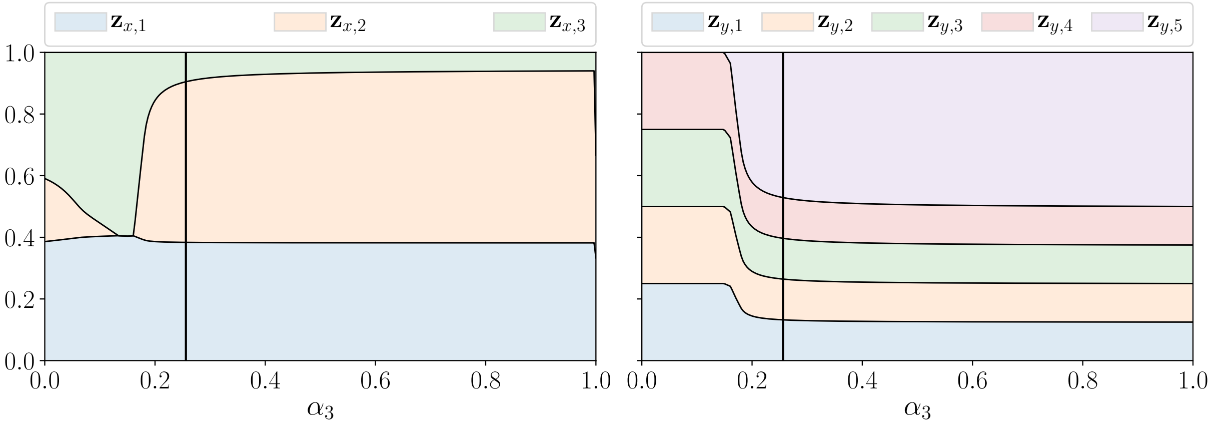

Figure 1: Feature importances $\mathbf{z}_x$ and $\mathbf{z}_y$ w.r.t. α_3 for the considered example

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

$$\alpha_1 = \frac{(1-\alpha_3)\overline{\mathbf{B}}}{\overline{\mathbf{Q}}_x + \overline{\mathbf{B}}}; \quad \alpha_2 = \frac{(1-\alpha_3)\overline{\mathbf{Q}}_x}{\overline{\mathbf{Q}}_x + \overline{\mathbf{B}}}; \quad \alpha_3 \in [0, 1]. \tag{10}$$

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 $\text{corr}(\boldsymbol{\nu}_1, \boldsymbol{\nu}_2) = 0.2$ and all others entries to one. Figure 1 shows the importances of features $\mathbf{z}_x$ and targets $\mathbf{z}_y$ with respect to α_3 coefficient. If α_3 is small, the impact of all targets are almost identical and the feature $\boldsymbol{\chi}_3$ dominates the feature $\boldsymbol{\chi}_2$. When α_3 becomes larger than 0.2, the importance $\mathbf{z}_{y,5}$ of the target $\boldsymbol{\nu}_5$ grows up along with the importance of the feature $\boldsymbol{\chi}_2$.

3.5 Minimax QPFS (MinMax and MaxMin).

The function (8) is symmetric with respect to $\mathbf{z}_x$ and $\mathbf{z}_y$. It penalizes the features that are correlated and are not relevant to the targets. At the same time it penalizes the targets that are correlated and are not sufficiently explained by the features. It leads to small importances for the targets which are difficult to predict by the features and large importances for the targets which are strongly correlated with the 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 the features. We express this into two related problems:

$$\alpha_1 \cdot \underbrace{\mathbf{z}_x^{\mathsf{T}}\mathbf{Q}_x\mathbf{z}_x}_{\text{Sim}(\mathbf{X})} - \alpha_2 \cdot \underbrace{\mathbf{z}_x^{\mathsf{T}}\mathbf{B}\mathbf{z}_y}_{\text{Rel}(\mathbf{X},\mathbf{Y})} \to \min_{\substack{\mathbf{z}_x \geq \mathbf{0}_n,\\ \mathbf{1}_n^{\mathsf{T}}\mathbf{z}_x = 1}}; \tag{11}$$

$$\alpha_3 \cdot \underbrace{\mathbf{z}_y^{\mathsf{T}}\mathbf{Q}_y\mathbf{z}_y}_{\text{Sim}(\mathbf{Y})} + \alpha_2 \cdot \underbrace{\mathbf{z}_x^{\mathsf{T}}\mathbf{B}\mathbf{z}_y}_{\text{Rel}(\mathbf{X},\mathbf{Y})} \to \min_{\substack{\mathbf{z}_y \geq \mathbf{0}_r,\\ \mathbf{1}_r^{\mathsf{T}}\mathbf{z}_y = 1}}. \tag{12}$$

The difference between (11) and (12) is the sign of Rel. In feature space the non-relevant components should have smaller importances. Meanwhile, the targets that are not relevant to the features should have larger importances. The problems (11) and (12) are merged into the joint min-max or max-min formulation

$$\min_{\substack{\mathbf{z}_x \geq \mathbf{0}_n \\ \mathbf{1}_n^\mathsf{T}\mathbf{z}_x = 1}} \max_{\substack{\mathbf{z}_y \geq \mathbf{0}_r \\ \mathbf{1}_r^\mathsf{T}\mathbf{z}_y = 1}} f(\mathbf{z}_x, \mathbf{z}_y), \quad \left(\text{or} \max_{\substack{\mathbf{z}_y \geq \mathbf{0}_r \\ \mathbf{1}_r^\mathsf{T}\mathbf{z}_y = 1}} \min_{\substack{\mathbf{z}_x \geq \mathbf{0}_n \\ \mathbf{1}_n^\mathsf{T}\mathbf{z}_x = 1}} f(\mathbf{z}_x, \mathbf{z}_y)\right), \tag{13}$$

where

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

Theorem 1. *For positive definite matrices $\mathbf{Q}_x$ and $\mathbf{Q}_y$ the max-min and min-max problems* (13) *have the same optimal value.*

Proof. Denote

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

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. $f(\cdot, \mathbf{z}_y) : \mathbb{C}^n \to \mathbb{R}$ is convex for fixed $\mathbf{z}_y$, and $f(\mathbf{z}_x, \cdot) : \mathbb{C}^r \to \mathbb{R}$ is concave for fixed $\mathbf{z}_x$. In this case Neumann's minimax theorem states

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

□

To solve the min-max problem (13), fix some $\mathbf{z}_x \in \mathbb{C}^n$. For fixed vector $\mathbf{z}_x$ we solve the problem

$$\max_{\mathbf{z}_y \in \mathbb{C}_r} f(\mathbf{z}_x, \mathbf{z}_y) = \max_{\substack{\mathbf{z}_y \geq \mathbf{0}_r \\ \mathbf{1}_r^\mathsf{T}\mathbf{z}_y = 1}} \left[\alpha_1 \cdot \mathbf{z}_x^\mathsf{T}\mathbf{Q}_x\mathbf{z}_x - \alpha_2 \cdot \mathbf{z}_x^\mathsf{T}\mathbf{B}\mathbf{z}_y - \alpha_3 \cdot \mathbf{z}_y^\mathsf{T}\mathbf{Q}_y\mathbf{z}_y\right]. \tag{14}$$

The Lagrangian for this problem is

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

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

$$\min_{\lambda,\, \boldsymbol{\mu} \geq \mathbf{0}_r} g(\mathbf{z}_x, \lambda, \boldsymbol{\mu}) = \min_{\lambda,\, \boldsymbol{\mu} \geq \mathbf{0}_r} \left[\max_{\mathbf{z}_y \in \mathbb{R}^r} L(\mathbf{z}_x, \mathbf{z}_y, \lambda, \boldsymbol{\mu})\right]. \tag{15}$$

The strong duality holds for quadratic problem (14) with positive definite matrices $\mathbf{Q}_x$ and $\mathbf{Q}_y$. Therefore, the optimal value for (14) equals the optimal value for (15). It allows to solve the problem

$$\min_{\mathbf{z}_x \in \mathbb{C}^n,\, \lambda,\, \boldsymbol{\mu} \geq \mathbf{0}_r} g(\mathbf{z}_y, \lambda, \boldsymbol{\mu}) \tag{16}$$

instead of (13).

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

$$\mathbf{z}_y = \frac{1}{2\alpha_3} \mathbf{Q}_y^{-1} \left(-\alpha_2 \cdot \mathbf{B}^{\mathsf{T}} \mathbf{z}_x + \lambda \cdot \mathbf{1}_r + \boldsymbol{\mu} \right). \tag{17}$$

The dual function is equal to

$$\begin{aligned} g(\mathbf{z}_x, \lambda, \boldsymbol{\mu}) = \max_{\mathbf{z}_y \in \mathbb{R}^r} L(\mathbf{z}_x, \mathbf{z}_y, \lambda, \boldsymbol{\mu}) = \mathbf{z}_x^{\mathsf{T}} \left(-\frac{\alpha_2^2}{4\alpha_3} \cdot \mathbf{B} \mathbf{Q}_y^{-1} \mathbf{B}^{\mathsf{T}} - \alpha_1 \cdot \mathbf{Q}_x \right) \mathbf{z}_x \\ - \frac{1}{4\alpha_3} \lambda^2 \cdot \mathbf{1}_r^{\mathsf{T}} \mathbf{Q}_y^{-1} \mathbf{1}_r - \frac{1}{4\alpha_3} \cdot \boldsymbol{\mu}^{\mathsf{T}} \mathbf{Q}_y^{-1} \boldsymbol{\mu} + \frac{\alpha_2}{2\alpha_3} \lambda \cdot \mathbf{1}_r^{\mathsf{T}} \mathbf{Q}_y^{-1} \mathbf{B}^{\mathsf{T}} \mathbf{z}_x \\ - \frac{1}{2\alpha_3} \lambda \cdot \mathbf{1}_r^{\mathsf{T}} \mathbf{Q}_y^{-1} \boldsymbol{\mu} + \frac{\alpha_2}{2\alpha_3} \cdot \boldsymbol{\mu}^{\mathsf{T}} \mathbf{Q}_y^{-1} \mathbf{B}^{\mathsf{T}} \mathbf{z}_x + \lambda. \end{aligned} \tag{18}$$

It brings to the quadratic problem (16) with $n + r + 1$ variables.

3.6 Asymmetric Importance (AsymImp)

Another way to overcome the problem of SymImp strategy is to add penalty for targets, which are well-explained by the features. We add the term $\mathbf{b}^{\mathsf{T}} \mathbf{z}_y$ to the term $\text{Rel}(\mathbf{X}, \mathbf{Y})$:

$$\alpha_1 \cdot \underbrace{\mathbf{z}_x^{\mathsf{T}} \mathbf{Q}_x \mathbf{z}_x}_{\text{Sim}(\mathbf{X})} - \alpha_2 \cdot \underbrace{\left(\mathbf{z}_x^{\mathsf{T}} \mathbf{B} \mathbf{z}_y - \mathbf{b}^{\mathsf{T}} \mathbf{z}_y \right)}_{\text{Rel}(\mathbf{X}, \mathbf{Y})} + \alpha_3 \cdot \underbrace{\mathbf{z}_y^{\mathsf{T}} \mathbf{Q}_y \mathbf{z}_y}_{\text{Sim}(\mathbf{Y})} \to \min_{\substack{\mathbf{z}_x \geq \mathbf{0}_n,\, \mathbf{1}_n^{\mathsf{T}} \mathbf{z}_x = 1 \\ \mathbf{z}_y \geq \mathbf{0}_r,\, \mathbf{1}_r^{\mathsf{T}} \mathbf{z}_y = 1}}. \tag{19}$$

Proposition 2. *Let the vector* $\mathbf{b}$ *equal*

$$b_j = \max_{i=1,\dots n} [\mathbf{B}]_{i,j}.$$

Then the importances coefficients for the vector $\mathbf{z}_y$ *will be nonnegative in term Rel*$(\mathbf{X}, \mathbf{Y})$ *for the problem* (19).

Proof. The proposition follows from the fact

$$\sum_{i=1}^{n} z_i b_{ij} \leq \left(\sum_{i=1}^{n} z_i \right) \max_{i=1,\dots n} b_{ij} = \max_{i=1,\dots n} b_{ij},$$

where $z_i \geq 0$ and $\sum_{i=1}^{n} z_i = 1$. □

Hence, the function (19) encourages the features which are relevant to the targets and encourages the targets that are not sufficiently correlated with the features.

Proposition 3. *The balance between the terms Sim*($\mathbf{X}$)*, Rel*($\mathbf{X}, \mathbf{Y}$)*, and Rel*($\mathbf{X}, \mathbf{Y}$) *for the problem* (19) *is achieved by the following coefficients:*

$$\alpha_1 \propto \overline{\mathbf{Q}}_y \left(\overline{\mathbf{b}} - \overline{\mathbf{B}}\right); \quad \alpha_2 \propto \overline{\mathbf{Q}}_x \overline{\mathbf{Q}}_y; \quad \alpha_3 \propto \overline{\mathbf{Q}}_x \overline{\mathbf{B}}.$$

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

$$\alpha_1 + \alpha_2 + \alpha_3 = 1; \tag{20}$$

$$\alpha_1 \overline{\mathbf{Q}}_x = \alpha_2 \overline{\mathbf{B}}; \tag{21}$$

$$\alpha_2 \left(\overline{\mathbf{b}} - \overline{\mathbf{B}}\right) = \alpha_3 \overline{\mathbf{Q}}_y. \tag{22}$$

Here we balance Sim($\mathbf{X}$) with the first term of Rel($\mathbf{X}, \mathbf{Y}$) by (21) and Sim($\mathbf{Y}$) with the full Rel($\mathbf{X}, \mathbf{Y}$) by (22). □

Proposition 4. *For the case $r = 1$ the proposed functions* (8)*,* (13)*,* (**??**)*, and* (19) *coincide with the original QPFS algorithm* (1)*.*

Proof. If r is equal to 1, then $\mathbf{Q}_y = q_y$ is a scalar, $\mathbf{z}_y = 1$, $\mathbf{B} = \mathbf{b}$. It reduces the problems (8), (13), and (19) to

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

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. RelAgg is the baseline strategy, which does not consider the target space correlations. SymImp penalizes the pairwise target correlations. MinMax more sensitive to the targets which are difficult for prediction. AsymImp strategy add the term to the SymImp function to make the features and targets influence asymmetric. The ideas in MinMax and AsymImp approaches are the same.

3.7 Dimensionality reduction

To eliminate the linear dependence and reduce the dimensionality of the input space, the principal components analysis (PCA) is widely used algorithm. The main disadvantage of the PCA method is its insensitivity to the interrelation between the features and the targets. The partial least squares algorithm projects the design matrix $\mathbf{X}$ and the target matrix $\mathbf{Y}$ to the latent space with low dimensionality ($l < n$). The PLS algorithm finds the latent space matrices $\mathbf{T}, \mathbf{U} \in \mathbb{R}^{m \times l}$ that best describe the original matrices $\mathbf{X}$ and $\mathbf{Y}$.

The design matrix $\mathbf{X}$ and the target matrix $\mathbf{Y}$ are projected into the latent space in the following way:

$$\underset{m \times n}{\mathbf{X}} = \underset{m \times l}{\mathbf{T}} \cdot \underset{l \times n}{\mathbf{P}^{\mathsf{T}}} + \underset{m \times n}{\mathbf{F}} = \sum_{k=1}^{l} \underset{m \times 1}{\mathbf{t}_k} \cdot \underset{1 \times n}{\mathbf{p}_k^{\mathsf{T}}} + \underset{m \times n}{\mathbf{F}}, \tag{23}$$

$$\underset{m \times r}{\mathbf{Y}} = \underset{m \times l}{\mathbf{U}} \cdot \underset{l \times r}{\mathbf{Q}^{\mathsf{T}}} + \underset{m \times r}{\mathbf{E}} = \sum_{k=1}^{l} \underset{m \times 1}{\mathbf{u}_k} \cdot \underset{1 \times r}{\mathbf{q}_k^{\mathsf{T}}} + \underset{m \times r}{\mathbf{E}}, \tag{24}$$

Algorithm	Idea	Error function $S(\mathbf{a}\|\mathbf{X},\mathbf{Y})$
RelAgg	$\min\bigl[\text{Sim}(\mathbf{X}) - \text{Rel}(\mathbf{X},\mathbf{Y})\bigr]$	$\min\limits_{\mathbf{z}_x}\bigl[(1-\alpha)\cdot\mathbf{z}_x^{\mathsf{T}}\mathbf{Q}_x\mathbf{z}_x - \alpha\cdot\mathbf{z}_x^{\mathsf{T}}\mathbf{B}\mathbf{1}_r\bigr]$
SymImp	$\min\bigl[\text{Sim}(\mathbf{X}) - \text{Rel}(\mathbf{X},\mathbf{Y})$ $+ \text{Sim}(\mathbf{Y})\bigr]$	$\min\limits_{\mathbf{z}_x,\,\mathbf{z}_y}\bigl[\alpha_1\cdot\mathbf{z}_x^{\mathsf{T}}\mathbf{Q}_x\mathbf{z}_x - \alpha_2\cdot\mathbf{z}_x^{\mathsf{T}}\mathbf{B}\mathbf{z}_y + \alpha_3\cdot\mathbf{z}_y^{\mathsf{T}}\mathbf{Q}_y\mathbf{z}_y\bigr]$
MinMax	$\min\bigl[\text{Sim}(\mathbf{X}) - \text{Rel}(\mathbf{X},\mathbf{Y})\bigr]$ $\max\bigl[\text{Rel}(\mathbf{X},\mathbf{Y}) + \text{Sim}(\mathbf{Y})\bigr]$	$\min\limits_{\mathbf{z}_x}\max\limits_{\mathbf{z}_y}\bigl[\alpha_1\cdot\mathbf{z}_x^{\mathsf{T}}\mathbf{Q}_x\mathbf{z}_x - \alpha_2\cdot\mathbf{z}_x^{\mathsf{T}}\mathbf{B}\mathbf{z}_y - \alpha_3\cdot\mathbf{z}_y^{\mathsf{T}}\mathbf{Q}_y\mathbf{z}_y\bigr]$
AsymImp	$\min\bigl[\text{Sim}(\mathbf{X}) - \text{Rel}(\mathbf{X},\mathbf{Y})\bigr]$ $\max\bigl[\text{Rel}(\mathbf{X},\mathbf{Y}) + \text{Sim}(\mathbf{Y})\bigr]$	$\min\limits_{\mathbf{z}_x,\mathbf{z}_y}\bigl[\alpha_1\cdot\mathbf{z}_x^{\mathsf{T}}\mathbf{Q}_x\mathbf{z}_x - \alpha_2\cdot\bigl(\mathbf{z}_x^{\mathsf{T}}\mathbf{B}\mathbf{z}_y - \mathbf{b}^{\mathsf{T}}\mathbf{z}_y\bigr) + \alpha_3\cdot\mathbf{z}_y^{\mathsf{T}}\mathbf{Q}_y\mathbf{z}_y\bigr]$

Table 1: Overview of the proposed multivariate QPFS algorithms

where $\mathbf{T}$, $\mathbf{U}$ are scores matrices in the latent space; $\mathbf{P}$, $\mathbf{Q}$ are loading matrices; $\mathbf{E}$, $\mathbf{F}$ are residual matrices. PLS maximizes the linear relation between columns of matrices $\mathbf{T}$ and $\mathbf{U}$

$$\mathbf{U} \approx \mathbf{T}\mathbf{B}, \quad \mathbf{B} = \text{diag}(\beta_k), \quad \beta_k = \mathbf{u}_k^{\mathsf{T}}\mathbf{t}_k/(\mathbf{t}_k^{\mathsf{T}}\mathbf{t}_k).$$

We use the PLS algorithm as the dimensionality reduction algorithm in this research.

To obtain the model prediction and find the model parameters, multiply the both hand sides of (23) by the matrix $\mathbf{W}$. Since the residual matrix $\mathbf{E}$ rows are orthogonal to the columns of $\mathbf{W}$, we have

$$\mathbf{X}\mathbf{W} = \mathbf{T}\mathbf{P}^{\mathsf{T}}\mathbf{W}.$$

The linear transformation between objects in the input and latent spaces is the following

$$\mathbf{T} = \mathbf{X}\mathbf{W}^*, \quad \text{where } \mathbf{W}^* = \mathbf{W}(\mathbf{P}^{\mathsf{T}}\mathbf{W})^{-1}. \tag{25}$$

The matrix of the model parameters (2) could be found from equations (24), (25)

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathsf{T}} + \mathbf{E} \approx \mathbf{T}\mathbf{B}\mathbf{Q}^{\mathsf{T}} + \mathbf{E} = \mathbf{X}\mathbf{W}^*\mathbf{B}\mathbf{Q}^{\mathsf{T}} + \mathbf{E} = \mathbf{X}\boldsymbol{\Theta} + \mathbf{E}. \tag{26}$$

Thus, the model parameters (2) are equal to

$$\boldsymbol{\Theta} = \mathbf{W}(\mathbf{P}^{\mathsf{T}}\mathbf{W})^{-1}\mathbf{B}\mathbf{Q}^{\mathsf{T}}.$$

The final model (26) is a linear model which are low-dimensional in the latent space. It reduces the data redundancy and increases the model stability.

4 Experiment

To evaluate the selected feature subset we introduce criteria that estimate the quality of feature selection. We measure multicorrelation by mean value of miltiple correlation coefficient as follows

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

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

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

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

A smaller value of Stability indicates less multicollinearity in the matrix $\mathbf{X}$.

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

$$\text{sRMSE}(\mathbf{Y}, \widehat{\mathbf{Y}}_{\mathbf{a}}) = \sqrt{\frac{\text{MSE}(\mathbf{Y}, \widehat{\mathbf{Y}}_{\mathbf{a}})}{\text{MSE}(\mathbf{Y}, \overline{\mathbf{Y}})}} = \frac{\|\mathbf{Y} - \widehat{\mathbf{Y}}_{\mathbf{a}}\|_2}{\|\mathbf{Y} - \overline{\mathbf{Y}}\|_2}.$$

Here $\widehat{\mathbf{Y}}_{\mathbf{a}} = \mathbf{X}_{\mathbf{a}}\boldsymbol{\Theta}_{\mathbf{a}}^{\mathsf{T}}$ is a model prediction and $\overline{\mathbf{Y}}$ is a constant prediction obtained by averaging the targets across all objects. The error on the test set should be as minimal as possible.

Bayesian Information Criteria (BIC) is a trade-off between prediction quality and the size of selected subset $\|\mathbf{a}\|_0$:

$$\text{BIC} = m \ln \Big(\text{MSE}(\mathbf{Y}, \widehat{\mathbf{Y}}_{\mathbf{a}}) \Big) + \|\mathbf{a}\|_0 \cdot \ln m,$$

where $\|\mathbf{a}\|_0 = \#\{j : a_j \neq 0\} = \sum_{j=1}^{n} a_j$. The less value of BIC means the better feature subset.

4.1 Data

We carried out computational experiment with ECoG data from the NeuroTycho project [21]. ECoG 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 with Morlet mother wavelet. The procedure of extracting feature representation from the raw data are described in details in [22,23]. We unfold the data and feature description at each time moment has dimension equals to 32 (channels) × 27 (frequencies) = 864. Each object is the representation of local history time segment 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. The example of initial brain signals and corresponding hand outcome are shown in Figure 2.

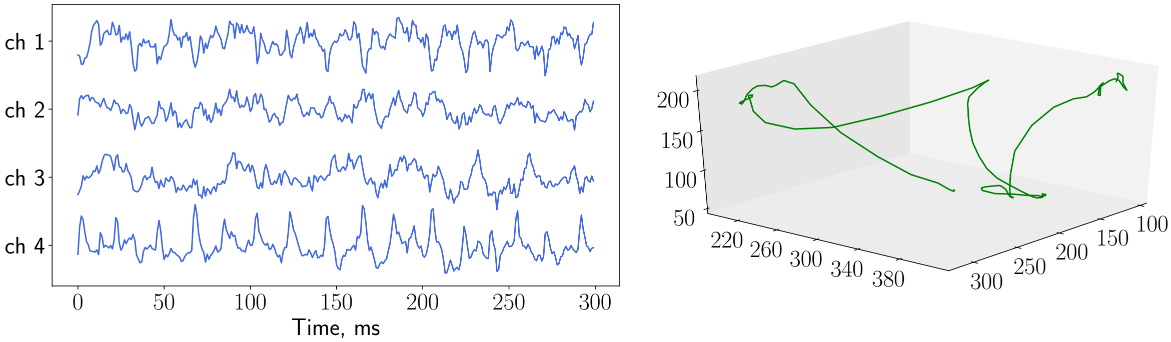

Figure 2: Brain signals (left plot) and 3D hand coordinates (right plot)

4.2 Results

Figure 3 shows the dependencies in the matrices $\mathbf{X}$ and $\mathbf{Y}$ for ECoG data. Frequencies in the matrix $\mathbf{X}$ are highly correlated. The frequencies are choosen in logarithmic scale, the closer the frequencies are the higher the correlations. In the target matrix $\mathbf{Y}$ the correlations between axes are not significant in comparison with the correlations between consequent moments and these correlations decay with time.

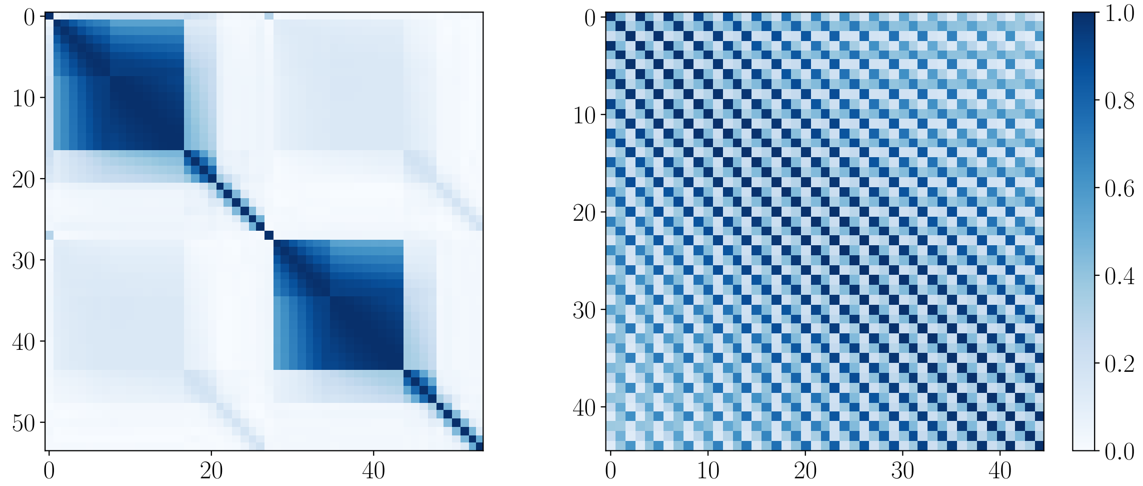

Figure 3: Correlation matrices for $\mathbf{X}$ and $\mathbf{Y}$

We apply the QPFS algorithm with SymImp strategy for different values of α_3 coefficient according to formulas (10). The dependence between target importances $\mathbf{z}_y$ with respect to α_3 for different values of k is shown in Figure 4. If we predict wrist coordinates only for one timestamp $k = 1$, targets importances are almost the same. It tells about the independence between x, y, and z coordinates. For $k = 2$ and $k = 3$ the importances of some targets become zero when α_3 increases. The vertical lines correspond to the optimal value of coefficient α_3 obtained by (9). The importances $\mathbf{z}_y$ for this value of α_3 are similar.

It means that the algorithm does not distinguish the targets for $k = 1, 2, 3$.

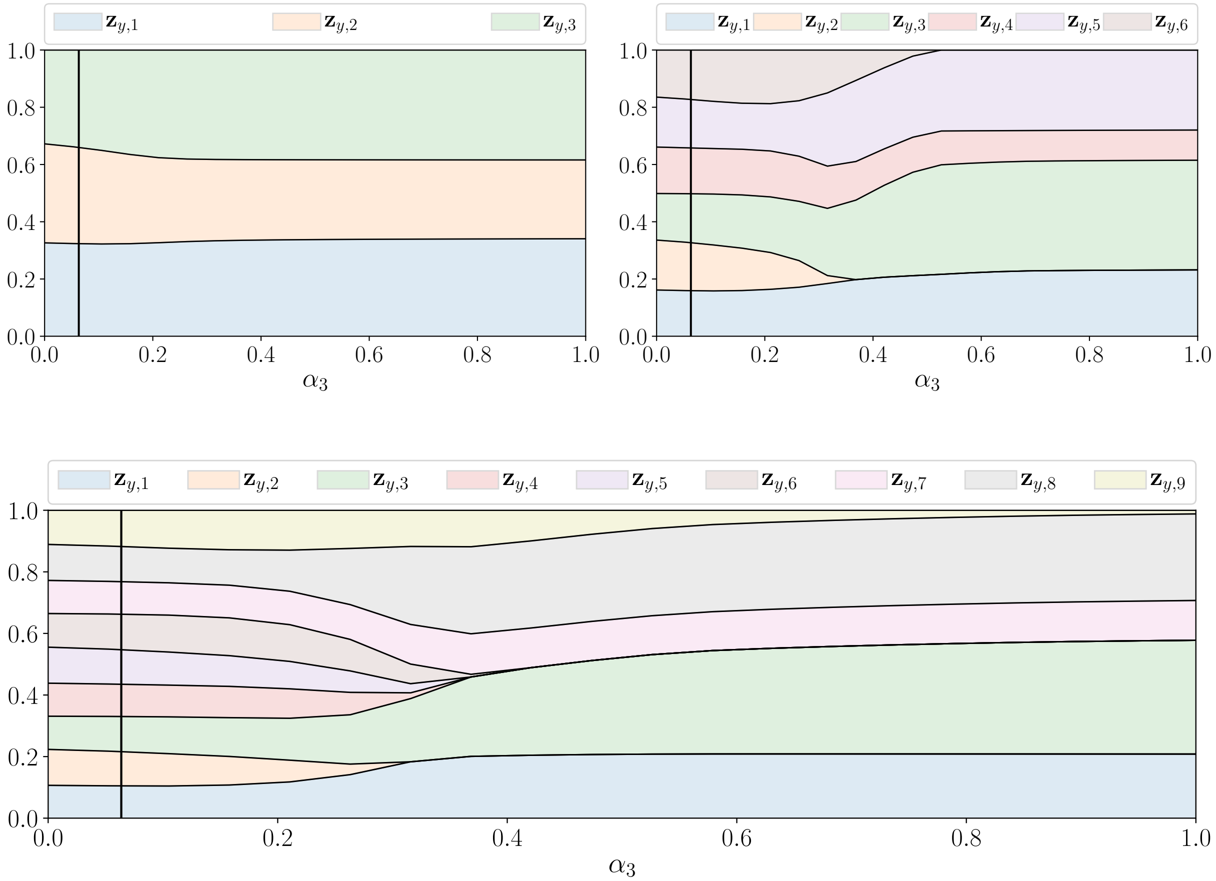

Figure 4: Target importances $\mathbf{z}_y$ with respect to α_3 for QPFS with Symmetric Importance

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 importances. Then we fit linear regression model with increasing number of features. For each method the features are sorted by the obtained importances. We show how the described metrics are changed with the increasing feature set size. Figure 5 illustrates the results for prediction of $k = 30$ timestamps. Here the feature importances threshold τ are shown by colored ticks. These thresholds are larger for the proposed methods with comparison to the baseline RelAgg strategy. The SymImp strategy has the largest threshold, it does not allow to get the small feature subset. However, this strategy shows the best performance in terms of sRMSE on test data. The second value of performance is given by AsymImp. All proposed algorithms give the less test error compared to the RelAgg strategy. The Stability criteria is also increased for the proposed algorithms. Here we consider the AsymImp strategy as the best in terms of prediction quality and the size of selected feature subset.

To compare the structure of the selected feature subsets and investigate the stability of the selection procedure, we use bootstrap approach. First, the bootstrap data are

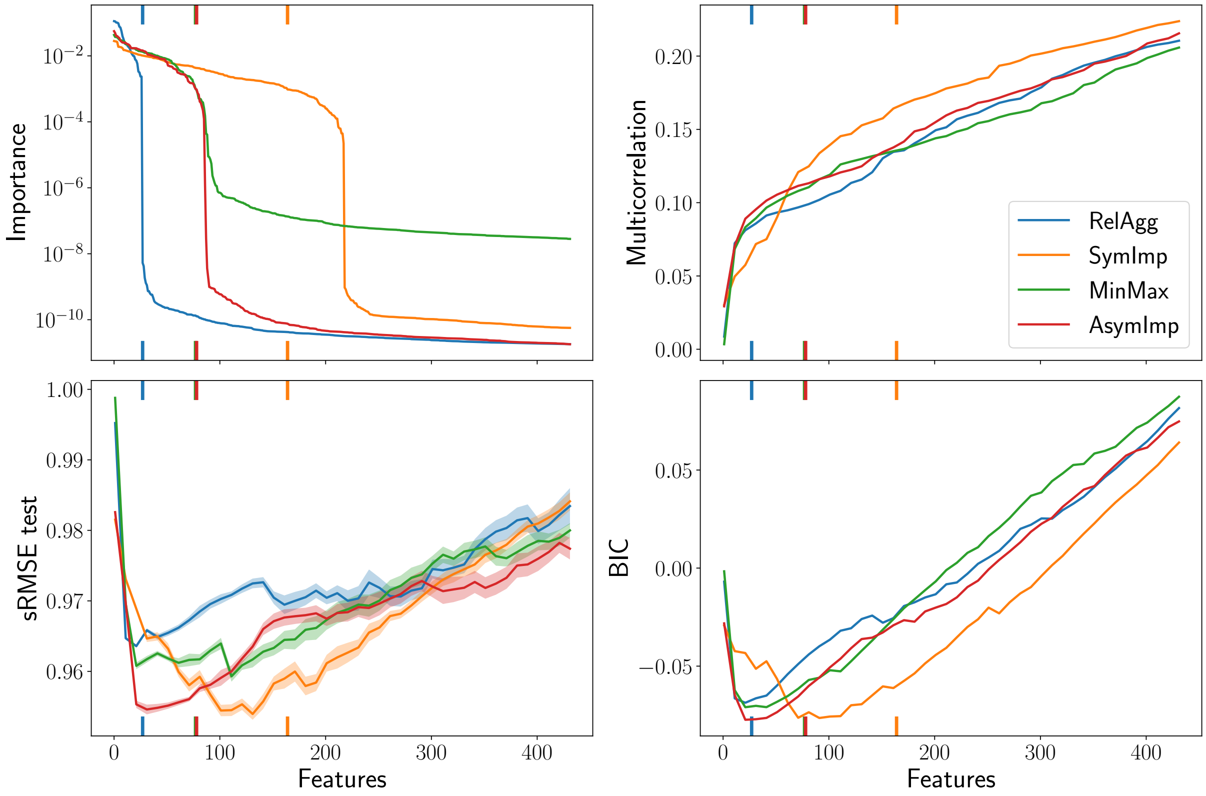

Figure 5: Feature selection algorithms evaluation for ECoG data, prediction of $k = 30$ timestamps

generated. Then solve the feature selection problem for each pair of the design and the target matrices. The obtained feature importances are compared. We calculate the average pairwise Spearman correlation coefficient and the ℓ_2 distance to obtain the measure of the algorithms stability. Table 2 shows the average error, the size of the subset and the described statistics for each method. The error was calculated by fitting the linear regression model on the 50 features with the largest importances. AsymImp gives the least error on the test data. The size of selected feature subsets are overestimated using the equal threshold $\tau = 10^{-4}$. The value of τ should be cross-validated to get the optimal threshold and the feature subset size.

We fit the PLS regression model for the data to compare the dimensionality reduction and feature selection. Figure 6 demonstrates the scaled RMSE on train and test data with respect to the dimensionality of the latent space l. The test error achieves minimum value at hte point $l = 11$. PLS regression is more flexible approach compared to the linear model built on the subset of features. It leads to the less error, but the model are not sparse.

Figure 7 compares 3 models: linear regression and PLS regression built on 100 features given by qpfs and PLS regression with all features. We do not include linear regression with all features because its results are close to the constant prediction. It also provides

Table 2: The stability of the selected feature subset

	sRMSE	$\|\mathbf{a}\|_0$	Spearman ρ	ℓ_2 dist
RelAgg	0.965 ± 0.002	26.8 ± 3.8	0.915 ± 0.016	0.145 ± 0.018
SymImp	0.961 ± 0.001	224.4 ± 9.0	0.910 ± 0.017	0.025 ± 0.002
MinMax	0.961 ± 0.002	101.0 ± 2.1	0.932 ± 0.009	0.059 ± 0.004
AsymImp	0.955 ± 0.001	85.8 ± 10.2	0.926 ± 0.011	0.078 ± 0.007

the result for Lasso and Elastic Net algorithms that are widely used for feature selection. We use the AsymImp strategy for QPFS in this experiment. The number of PLS latent dimension is $l = 15$. Here PLS regression are significantly better than linear regression with QPFS features. It means that the latter model is not flexible enough. However, the best result is obtained by combination of PLS regression model with QPFS features. This model is sparse since it uses only 100 QPFS features. The ability of the PLS model to find the optimal latent data representation allows to improve model performance.

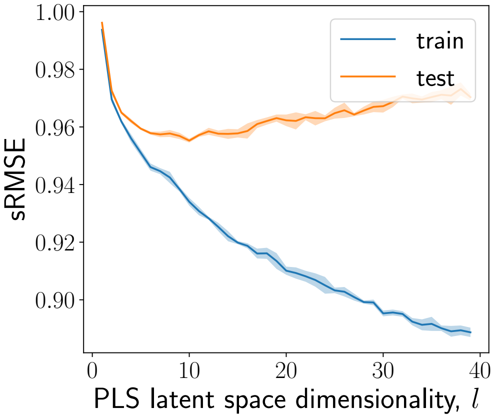

Figure 6: Test scaled RMSE for PLS regression model

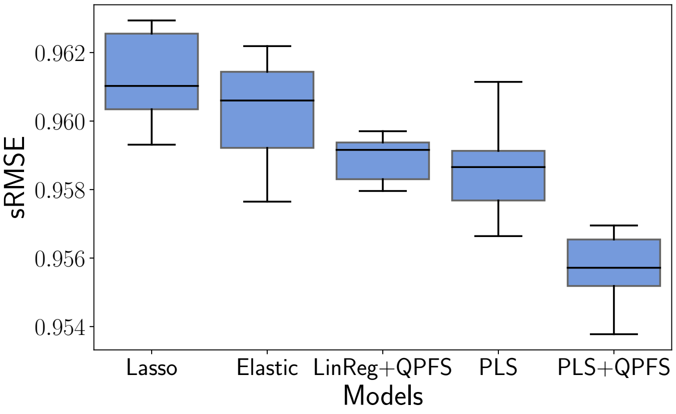

Figure 7: sRMSE box plots for different models

5 Conclusion

The study investigates the problem of signal decoding. The data are highly redundant. To build a stable edequate model, it was proposed to reduce dimensionality of the problem using the dependencies in both input and target spaces. The PLS regression is considered as linear model for dimensionality reduction. The quadratic programming approach is investigated as feature selection algorithm. The algorithm solves feature selection in a single quadratic programming optimization problem. The multivariate extensions for the QPFS algorithms are proposed. The resulting feature subset includes non-correlated features which are relevant to the most difficult targets.

The computational experiments were carried out on the ECoG data. The resulting model predicts the limb position of an exoskeleton by brain signals. The proposed algorithms outperforms the baseline algorithm and reduce the problem dimension significantly. The combination of feature selection for sparsifying the model and the dimensionality reduction for increasing model stability give the best result.

References

[1] Sadegh Karimi and Maryam Farrokhnia. Leukemia and small round blue-cell tumor cancer detection using microarray gene expression data set: Combining data dimension reduction and variable selection technique. *Chemometrics and Intelligent Laboratory Systems*, 139:6–14, 2014.

[2] You-Wu Lin, Bai-Chuan Deng, Qing-Song Xu, Yong-Huan Yun, and Yi-Zeng Liang. The equivalence of partial least squares and principal component regression in the sufficient dimension reduction framework. *Chemometrics and Intelligent Laboratory Systems*, 150:58–64, 2016.

[3] Andrey Eliseyev and Tatiana Aksenova. Stable and artifact-resistant decoding of 3d hand trajectories from ecog signals using the generalized additive model. *Journal of neural engineering*, 11(6):066005, 2014.

[4] Andrey Eliseyev, Cecile Moro, Jean Faber, Alexander Wyss, Napoleon Torres, Corinne Mestais, Alim Louis Benabid, and Tetiana Aksenova. L1-penalized n-way pls for subset of electrodes selection in bci experiments. *Journal of neural engineering*, 9(4):045010, 2012.

[5] Hyonho Chun and Sündüz Keleş. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):3–25, 2010.

[6] Tahir Mehmood, Kristian Hovde Liland, Lars Snipen, and Solve Sæbø. A review of variable selection methods in partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 118:62–69, 2012.

[7] A. M. Katrutsa and V. V. Strijov. Stress test procedure for feature selection algorithms. *Chemometrics and Intelligent Laboratory Systems*, 142:172–183, 2015.

[8] 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.

[9] Julien Lauzon-Gauthier, Petre Manolescu, and Carl Duchesne. The sequential multiblock pls algorithm (smb-pls): Comparison of performance and interpretability. *Chemometrics and Intelligent Laboratory Systems*, 2018.

[10] Sarah Engel, Tetiana Aksenova, and Andrey Eliseyev. Kernel-based npls for continuous trajectory decoding from ecog data for bci applications. In *International Conference on Latent Variable Analysis and Signal Separation*, pages 417–426. Springer, 2017.

[11] Alessandra Biancolillo, Tormod Næs, Rasmus Bro, and Ingrid Måge. Extension of so-pls to multi-way arrays: So-n-pls. *Chemometrics and Intelligent Laboratory Systems*, 164:113–126, 2017.

[12] D. Hervás, J.M. Prats-Montalbán, A. Lahoz, and A. Ferrer. Sparse n-way partial least squares with r package snpls. *Chemometrics and Intelligent Laboratory Systems*, 179:54 – 63, 2018.

[13] Roman Rosipal and Nicole Kramer. Overview and Recent Advances in Partial Least Squares. *C. Saunders et al. (Eds.): SLSFS 2005, LNCS 3940*, pages 34–51, 2006.

[14] Roman Rosipal. Nonlinear partial least squares an overview. In *Chemoinformatics and advanced machine learning perspectives: complex computational methods and collaborative techniques*, pages 169–189. IGI Global, 2011.

[15] 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.

[16] Makoto Yamada, Wittawat Jitkrittum, Leonid Sigal, Eric P Xing, and Masashi Sugiyama. High-dimensional feature selection by feature-wise kernelized lasso. *Neural computation*, 26(1):185–207, 2014.

[17] 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.

[18] 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.

[19] Anastasia Motrenko and Vadim Strijov. Multi-way feature selection for ecog-based brain-computer interface. *Expert Systems with Applications*, 2018.

[20] Yamuna Prasad, KK Biswas, and Parag Singla. Scaling-up quadratic programming feature selection. In *AAAI (Late-Breaking Developments)*, 2013.

[21] Kentaro Shimoda, Yasuo Nagasaka, Zenas C Chao, and Naotaka Fujii. Decoding continuous three-dimensional hand trajectories from epidural electrocorticographic signals in japanese macaques. *Journal of neural engineering*, 9(3):036015, 2012.

[22] 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.

[23] 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.