Quadratic Programming Feature Selection for Multicorrelated Signal Decoding with Partial Least Squares*

R.V. Isachenko[a,*], V.V. Strijov[a]

[a] *Moscow Institute of Physics and Technology, 9 Institutskiy Per., Dolgoprudny, Moscow Region 141700, Russian Federation*

Abstract

This paper investigates dimensionality reduction problem for signal decoding. Its main application is brain-computer interface modelling. The challenge is high redundancy in the data description. Data combines time series of two origins: design space: brain cortex signals and target space: limb motion signals. High correlations among measurements of complex signals lead to multiple correlations. This case studies correlations in both input and target spaces that carry heterogeneous data. This paper proposes feature selection algorithms to construct simple and stable forecasting model. It extends ideas of the quadratic programming feature selection approach and selects non-correlated features that are relevant to the target. The proposed methods take into account dependencies in both design and target space and select features, which fit both spaces jointly. The computational experiment was carried out using an electrocorticogram (ECoG) dataset. The obtained models predict hand motions using signals of the brain cortex. The partial least squares (PLS) regression model is used as the base model for dimensionality reduction. The best result is obtained by PLS algorithm, that reduces space dimensionality using the QPFS.

Keywords: partial least squares, quadratic programming feature selection, signal decoding, electrocorticogram

ere

1. Introduction

The raw data in the fields of chemometrics (Katrutsa and Strijov, 2015; Karimi and Farrokhnia, 2014; Lin et al., 2016) and signal decoding (Motrenko and Strijov, 2018; Eliseyev and Aksenova, 2014; Eliseyev et al., 2012) are high-dimensional and extremely redundant. The models built on such data are unstable. The redundant data description requires

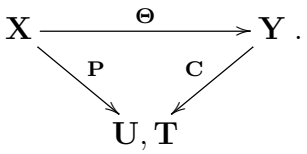
*The research was made possible by Government of the Russian Federation (Agreement 05.Y09.21.0018)

*Corresponding author

Email addresses: roman.isachenko@phystech.edu (R.V. Isachenko), strijov@phystech.edu (V.V. Strijov)

excessive computations, which lead to extended analysis time. To overcome this problem, dimensionality reduction (Chun and Keleş, 2010; Mehmood et al., 2020) and feature selection (Katrutsa and Strijov, 2015; Li et al., 2017) methods are used for high-dimensional data modeling.

The partial least squares (PLS) regression is a widely used algorithm for dimensionality reduction (Lauzon-Gauthier et al., 2018; Engel et al., 2017; Biancolillo et al., 2017; Hervás et al., 2018). The PLS projects initial features into low-dimensional space and uses the new feature description as model features. It maps the features and the targets onto joint latent space and maximizes the covariances between the projected vectors. It retrieves the information about the initial input and target matrices and extracts their relations. The following diagram shows the main principles of forecasting for the case, where both the source $\mathbf{X}$ and the target $\mathbf{Y}$ lie in spaces that have hidden dependencies, which could be reduced. The model projects input and target data into joint latent space and maximizes covariances between the projections



Parameters $\mathbf{\Theta}$ of the linear model are set in order to obtain a better forecast subject to the condition $\operatorname{cov}(\mathbf{U}, \mathbf{T}) \to \max$. The latent variables $\mathbf{U}, \mathbf{T}$ approximate hidden dependencies in joint space using linear models with parameters $\mathbf{P}, \mathbf{C}$.

An overview of the advances in the PLS regression is given in (Rosipal and Krämer, 2006; Rosipal, 2010). The dimensionality of latent space is much less than the size of the raw data description. The selected features makes the linear model more stable. In this case, we obtain the linear model with a small latent space dimensionality. However, the final model uses the whole range of the raw features, and it does not allow to remove useless features.

To obtain a sparse and stable forecasting model, we apply feature selection procedure. The majority of feature selection methods ignores dependencies in target space. This study suggests a novel approach to feature selection. The proposed approach extends the ideas of the quadratic programming feature selection (QPFS) (Rodriguez-Lujan et al., 2010) algorithm. It selects the non-correlated features that are relevant to the targets. The proposed methods take into account the dependencies in target space and select features, which are jointly informative to all targets.

Feature selection is a special type of dimensionality reduction where the latent representation is a subset of the initial data description. Here, a subset of features defines a forecasting model. The QPFS maximizes the feature relevances and minimizes the pairwise feature redundancy. This approach was proposed and investigated in (Ding and Peng, 2005; Yamada et al., 2014). It solves the optimization problem to select an optimal feature set. The paper (Katrutsa and Strijov, 2017) shows that the QPFS algorithm outperforms many existing feature selection methods for the univariate regression problem. The admissible

feature set is denoted in (1) by the vector $\mathbf{z}$. Each selected feature is indicated by one element of this vector, which makes 2^n-1 admissible combinations of features. The solution $\mathbf{z}$ of the quadratic form (1) is delivered by a convex optimization algorithm with relaxation of the integer variable, here $\mathbf{z} \in \mathbb{R}^n$. The algorithm solves the following optimization problem:

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

The QPFS algorithm introduces two functions: Sim and Rel. Sim estimates the redundancy between features, and Rel renders the relevances between each feature and the target vector. The QPFS minimizes the Sim function and maximizes the Rel function simultaneously. The columns of matrix $\mathbf{X}$ are the features, and $\mathbf{Y}$ is the target. The entries of matrix $\mathbf{Q} \in \mathbb{R}^{n\times n}$ 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 $\mathbf{z} \in \mathbb{R}^n$ shows the importance of each feature. Function (1) penalizes the dependent features using the Sim function and promotes the features that are relevant to the target using the Rel function. The parameter α controls the trade-off between Sim and Rel. To measure the similarity, the authors of (Rodriguez-Lujan et al., 2010) used the absolute value of the sample correlation coefficient between pairs of features for the Sim function, and between the features and the target vector for the Rel function. The detailed explanation of this optimization problem is given in next section.

Paper (Motrenko and Strijov, 2018) proposes a multi-way version of the QPFS algorithm for ECoG-based tensor data. It was shown that QPFS is an appropriate feature selection method for the signal decoding problem. This paper investigates the multivariate problem, where the dependent variable is a vector. It leads to correlations in the targets. In this case, feature selection algorithms do not take into account these dependencies. Hence, the selected feature subset is not optimal regarding its prediction.

We propose methods that take into account the dependencies in both input and target spaces. It allows to form a stable sparse model. We refer to the original QPFS algorithm as our baseline for the computational experiment. The convergence of the quadratic programming algorithms were shown theoretically in (Isachenko and Strijov, 2018). The class of convex algorithms was investigated by (Nesterov, 1983) and (Blaschke, 1996).

The main drawback of the QPFS algorithm is its computational costs. However, the original paper (Rodriguez-Lujan et al., 2010) suggests a way to solve the quadratic problem (1) efficiently. Additionally, in (Prasad et al., 2013) the proposed sequential minimal optimization framework solves the problem (1). There are alternative ways for solving such optimization problems (Preitl et al., 2006; Chiang et al., 2014; Li et al., 2020; Precup et al., 2021). Our choice of the QPFS formulation is based on the applicability of this approach to the ECOG data (Motrenko and Strijov, 2018), its success on the feature selection task (Katrutsa and Strijov, 2017), and the interpretability of the solution in terms of linear multicorrelation.

The experiments were carried out using the ECoG dataset (Shimoda et al., 2012). We compared the proposed methods for multivariate feature selection with the baseline

strategy and the PLS algorithm (Isachenko et al., 2018). The stability of the proposed methods was investigated by measuring how the feature selection solution changes with data bootstrapping. Given the same number of features the proposed algorithms outperform the baseline algorithm. The combination of the feature selection procedure and the PLS algorithm gives the best performance.

The main contributions of this paper are:

1. Addressing the dimensionality reduction problem for high-dimensional input and target data;
2. Proposing new feature selection methods for multivariate regression with the analysis of input and target spaces structures;
3. Comparing the proposed methods using a real ECoG dataset, and showing that the proposed methods give better feature subsets than the baseline method.

2. Multivariate regression problem

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 that there is a linear dependence between the object $\mathbf{x}$ and the target variable $\mathbf{y}$ as

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

where $\boldsymbol{\Theta} \in \mathbb{R}^{r \times n}$ is the matrix of the model parameters, and $\boldsymbol{\varepsilon} \in \mathbb{R}^r$ 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 and $\mathbf{Y} \in \mathbb{R}^{m \times r}$ 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 $\mathbf{X}$ correspond to the object features, and the columns $\boldsymbol{\nu}_j$ of $\mathbf{Y}$ correspond to the targets.

The optimal parameters are determined by the minimization of an error function. We define the quadratic loss function as follows:

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

The solution of (3) is given by

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

The linear dependent columns of $\mathbf{X}$ lead to an unstable 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 is not invertible. To avoid strong linear dependence, dimensionality reduction and feature selection are used.

3. Feature selection problem

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

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

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 Table 1.

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

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

Here, the vector $\mathbf{z}$ entries are the normalized feature importances. First, we solve problem (5) to obtain the feature importances $\mathbf{z}$. Then, the solution of (4) is recovered by setting a threshold as follows:

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

τ is a hyperparameter that is defined manually or chosen by cross-validation. The problem (5) is the main problem for feature selection task used in the paper.

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

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

where subscript $\mathbf{a}$ indicates the sub matrix with the columns in which the components of $\mathbf{a}$ equal 1. Further in this section we derive the proposed feature selection methods.

3.1. Quadratic Programming Feature Selection

Paper (Katrutsa and Strijov, 2017) shows that the QPFS outperforms many existing feature selection algorithms using different quality criteria. The QPFS algorithm selects the non-correlated features that are relevant to the target vector $\boldsymbol{\nu}$ for the linear regression problem where $r = 1$ as follows:

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

The authors of the original QPFS paper (Rodriguez-Lujan et al., 2010) suggested the following way to select α for (1) and make $\mathrm{Sim}(\mathbf{X})$ and $\mathrm{Rel}(\mathbf{X}, \boldsymbol{\nu})$ have the same impacts:

$$\alpha = \frac{\overline{\mathbf{Q}}}{\overline{\mathbf{Q}} + \overline{\mathbf{b}}}, \quad \overline{\mathbf{Q}} = \mathrm{mean}(\mathbf{Q}), \;\; \overline{\mathbf{b}} = \mathrm{mean}(\mathbf{b}).$$

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 the sample Pearson correlation coefficient:

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

Other ways to define $\mathbf{Q}$ and $\mathbf{b}$ are considered in (Katrutsa and Strijov, 2017).

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

3.2. Multivariate QPFS

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

Relevance aggregation (RelAgg). In (Motrenko and Strijov, 2018), in order to apply the QPFS algorithm to the multivariate case ($r > 1$), feature relevances are aggregated through all r components. The term $\mathrm{Sim}(\mathbf{X})$ is still the same, and matrix $\mathbf{Q}$ is defined by (6). The vector $\mathbf{b}$ is aggregated across all targets and is defined as

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

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

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

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

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

We would like to select only two features. For such a configuration, the best feature subset

is $[\boldsymbol{\chi}_1, \boldsymbol{\chi}_2]$. Feature $\boldsymbol{\chi}_2$ predicts the second target $\boldsymbol{\nu}_2$ and feature combination $\boldsymbol{\chi}_1, \boldsymbol{\chi}_2$ 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 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 extensions to the QPFS algorithm that overcome this example challenge.

Symmetric importances (SymImp). To take into account the dependencies in the columns of 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:

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

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

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

Vector $\mathbf{z}_x$ shows the features' importances, while $\mathbf{z}_y$ is a vector of the targets importances. The correlated targets will be penalized by $\mathrm{Sim}(\mathbf{Y})$ and have lower importances.

The coefficients α_1, α_2, and α_3 control the influence of each term on function (8) and satisfy the following 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:*

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

Proof. The desired values of α_1, α_2, and α_3 are given by solving 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 $\overline{\mathbf{Q}}_x$, $\overline{\mathbf{B}}$, 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 $\mathrm{Sim}(\mathbf{Y})$. □

To investigate the impact of $\mathrm{Sim}(\mathbf{Y})$ on function (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:

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

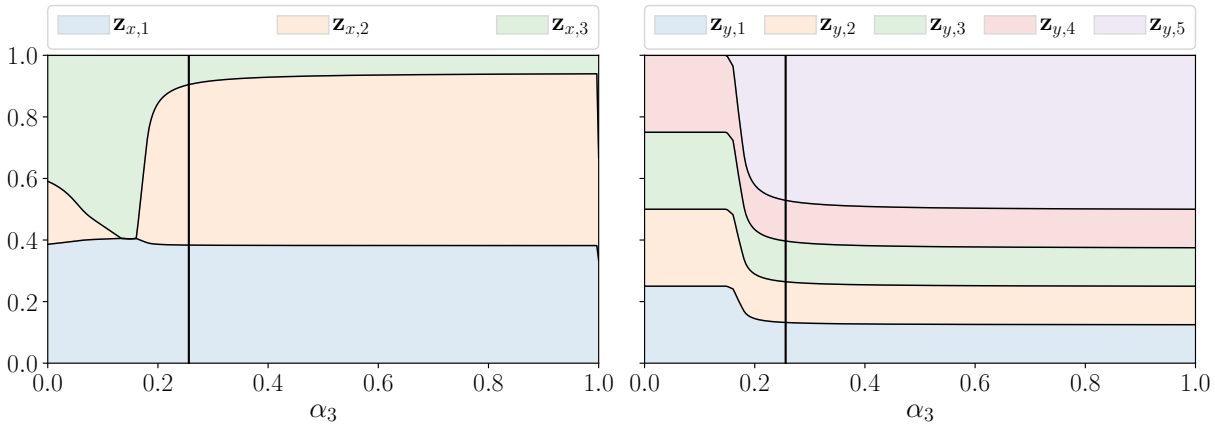


Figure 1: Feature importances $\mathbf{z}_x$ and $\mathbf{z}_y$ with respect to α_3 for the considered example

We apply the proposed algorithm to the discussed example (7). The given matrix $\mathbf{Q}$ corresponds to matrix $\mathbf{Q}_x$. We additionally define matrix $\mathbf{Q}_y$ by setting $\operatorname{corr}(\boldsymbol{\nu}_1, \boldsymbol{\nu}_2) = 0.2$ 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. If α_3 is small, the impacts of all targets are almost identical and feature $\boldsymbol{\chi}_3$ dominates feature $\boldsymbol{\chi}_2$. When α_3 becomes larger than 0.2, the importance $\mathbf{z}_{y,5}$ of target $\boldsymbol{\nu}_5$ increases along with the importance of feature $\boldsymbol{\chi}_2$.

Minimax QPFS (MinMax). Function (8) is symmetric with respect to $\mathbf{z}_x$ and $\mathbf{z}_y$. It penalizes the features that are correlated and irrelevant to the targets. In addition, it penalizes the targets that are correlated and are not sufficiently explained by the features. It leads to small importances for the targets that are weakly correlated with the features and large importances for the targets that are strongly correlated with the features. This result contradicts the intuition. Our goal is to predict all targets, especially those that are difficult to explain, using the selected relevant and non-correlated features. We express this as two related problems:

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

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

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

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

where

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

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. We denote the following:

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

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

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

□

To solve the min-max problem (13), we fix some $\mathbf{z}_x \in \mathbb{C}^n$. For a 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 $\boldsymbol{\mu}$ that correspond to the inequality constraints $\mathbf{z}_y \geq \mathbf{0}_r$ are restricted to being 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 the positive definite matrices $\mathbf{Q}_x$ and $\mathbf{Q}_y$. Therefore, the optimal value for (14) equals the optimal value for (15). It allows us 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).

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

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

The dual function is equal to

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

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

Asymmetric Importance (AsymImp). The natural way to overcome the problem of the SymImp strategy is to add penalties for targets that are correlated with features. We add the term $\mathbf{b}^\mathsf{T}\mathbf{z}_y$ to the term $\text{Rel}(\mathbf{X}, \mathbf{Y})$ as follows:

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

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

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

Then, the importance coefficients for vector $\mathbf{z}_y$ *will be non-negative in Rel*$(\mathbf{X}, \mathbf{Y})$ *for problem* (19)*.*

Proof. The proposition follows from the fact that

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

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

Hence, function (19) encourages the features that 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 the solutions to the following equations:

$$\alpha_1 + \alpha_2 + \alpha_3 = 1; \quad (20)$$
$$\alpha_1 \overline{\mathbf{Q}}_x = \alpha_2 \overline{\mathbf{B}}; \quad (21)$$
$$\alpha_2 \left(\overline{\mathbf{b}} - \overline{\mathbf{B}} \right) = \alpha_3 \overline{\mathbf{Q}}_y. \quad (22)$$

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

Table 1: Overview of the proposed multivariate QPFS algorithms

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

Proposition 4. *For the case of $r = 1$, 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, $\mathbf{z}_y = 1$, and $\mathbf{B} = \mathbf{b}$. It reduces 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 $\alpha = \frac{\alpha_2}{\alpha_1+\alpha_2}$ represents the original QPFS problem (1). □

Table 1 shows the core ideas and error functions for each method and summarizes all the proposed strategies for multivariate feature selection. RelAgg is the baseline strategy, and it does not consider target space correlations. SymImp penalizes the pairwise target correlations. MinMax are more sensitive to the targets that are difficult to predict. The AsymImp strategy adds the term to the SymImp function to make the features and targets have asymmetric influences. All the proposed methods solve the quadratic programming optimization tasks. The convergence of these type of problems carefully studied in (Nesterov, 1983; Blaschke et al., 1997; Isachenko and Strijov, 2018).

3.3. Dimensionality reduction

To eliminate the linear dependence and reduce the dimensionality of input space, principal components analysis (PCA) is a widely used algorithm. The main disadvantage of the PCA method is that it is insensitive 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 latent space with low dimensionality ($l < n$). The PLS algorithm finds 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 latent space in the

following way:

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

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

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

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

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

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

$$\mathbf{XW} = \mathbf{TP}^{\mathsf{T}}\mathbf{W}.$$

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

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

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

$$\mathbf{Y} = \mathbf{UC}^{\mathsf{T}} + \mathbf{E} \approx \mathbf{TBC}^{\mathsf{T}} + \mathbf{E} = \mathbf{XW}^*\mathbf{BC}^{\mathsf{T}} + \mathbf{E} = \mathbf{X\Theta} + \mathbf{E}. \tag{26}$$

Thus, the model parameters (2) are equal to

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

The final model (26) is a linear model that are low-dimensional in 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 the multicorrelation using the mean value of multiple correlation coefficient as follows:

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

This coefficient lies between 0 and 1. A bigger R^2 means that we have a better feature subset.

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

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

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

The scaled Root Mean Squared Error (sRMSE) shows the quality of the model prediction. We estimate the sRMSE using 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, $\widehat{\mathbf{Y}}_\mathbf{a} = \mathbf{X}_\mathbf{a}\mathbf{\Theta}_\mathbf{a}^\mathsf{T}$ is the model prediction and $\overline{\mathbf{Y}}$ is the constant prediction obtained by averaging the targets across all objects. The error on the test set should be as small as possible.

The Bayesian Information Criteria (BIC) incorporates a trade-off between the prediction quality and the size of selected subset $\|\mathbf{a}\|_0 = \#\{j : a_j \neq 0\} = \sum_{j=1}^n a_j$:

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

A smaller value of BIC means a better feature subset.

All convex quadratic optimization problems had been solved using CVXPY open-source Python library (Diamond and Boyd, 2016; Agrawal et al., 2018). The source code for the paper could be found here [1].

4.1. Data

We conducted a computational experiment with the ECoG data from the NeuroTycho project (Shimoda et al., 2012; Motrenko and Strijov, 2018). This data and the problem of Brain-Computer Interface construction itself demanded the proposed theory due to the high multicorrelation and redundancy of the features. The paper (Motrenko and Strijov, 2018) shows experimentally three types of multicorrelated data. First, the ECoG signals are cross-correlated due the short distance between the cortical electrodes. Second, the limb movement sequence is interdependent due its physical nature. And third, there is the multiple correlation between the brain and limb signals, which established is the main reason to construct the BCI forecasting models.

The ECoG data consist of brain voltage signals recorded over 32 channels. The goal is to predict 3D hand positions in subsequent moments given the input signal. The initial voltage signals are transformed to the spatial-temporal representation using the wavelet transformation with the Morlet mother wavelet. The procedure of extracting the feature representation from the raw data is described in detail in (Chao et al., 2010; Eliseyev and

[1] https://github.com/Intelligent-Systems-Phystech/MultivariateQPFS

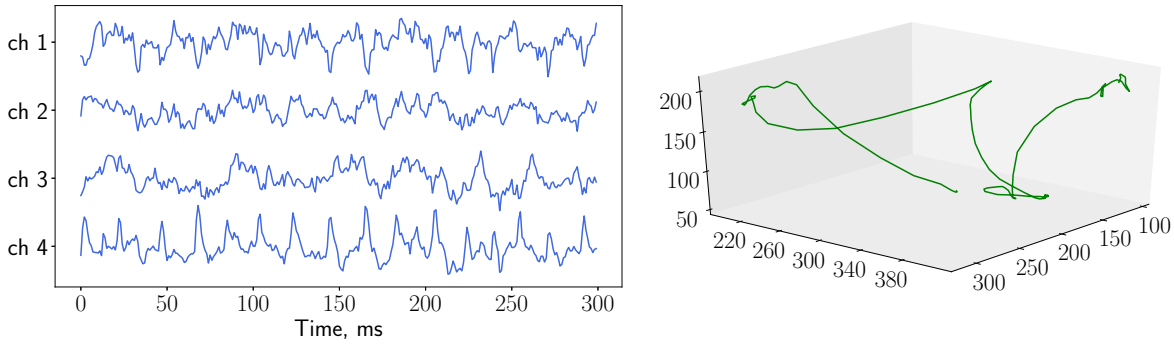


Figure 2: Brain signals (left plot) and 3D hand coordinates (right plot)

Aksenova, 2016). We unfold the data and feature description at each time moment has dimension of size 32 (channels) $\times$ 27 (frequencies) = 864. Each object is the representation of the local historical 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 the number of timestamps that we predict. We split our data into train and test parts with the ratio 0.67. Example of the initial brain signals and the corresponding hand trajectory is shown in Figure 2.

4.2. Results

Figure 3 shows the dependencies in the matrices $\mathbf{X}$ and $\mathbf{Y}$ for the ECoG data. The frequencies in the matrix $\mathbf{X}$ are highly correlated. 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.

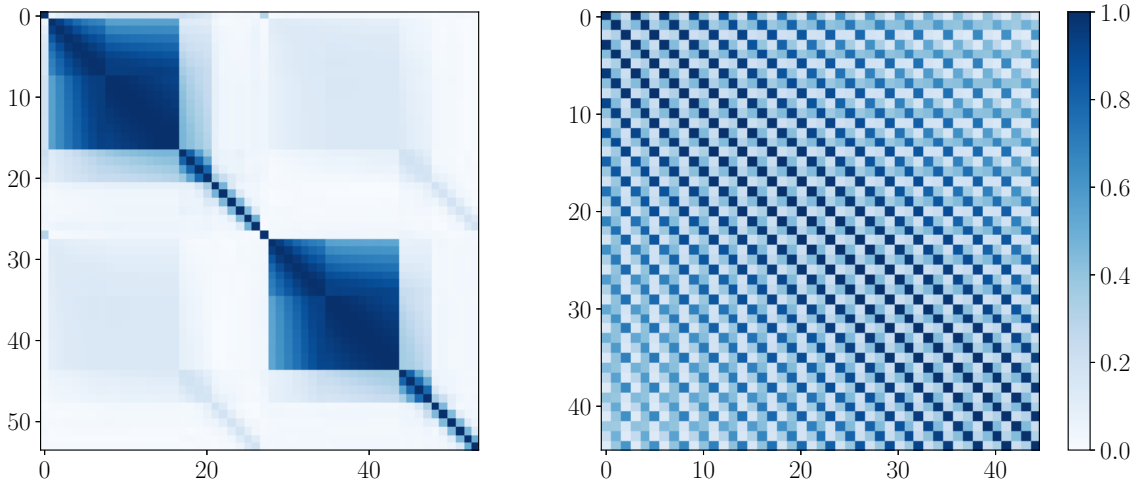


Figure 3: Correlation matrices for $\mathbf{X}$ and $\mathbf{Y}$

We apply the QPFS algorithm with the SymImp strategy for different values of α_3 according to formula (10). The dependencies between target importances $\mathbf{z}_y$ with respect

to α_3 for different values of k are shown in Figure 4. The targets importances are almost the same for the predicted wrist coordinates with only one timestamp $k = 1$, which reflects the independence between the 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 α_3 obtained by (9). The target importances $\mathbf{z}_y$ for this value of α_3 are similar. Thus, the algorithm does not distinguish the targets for $k = 1, 2, 3$.

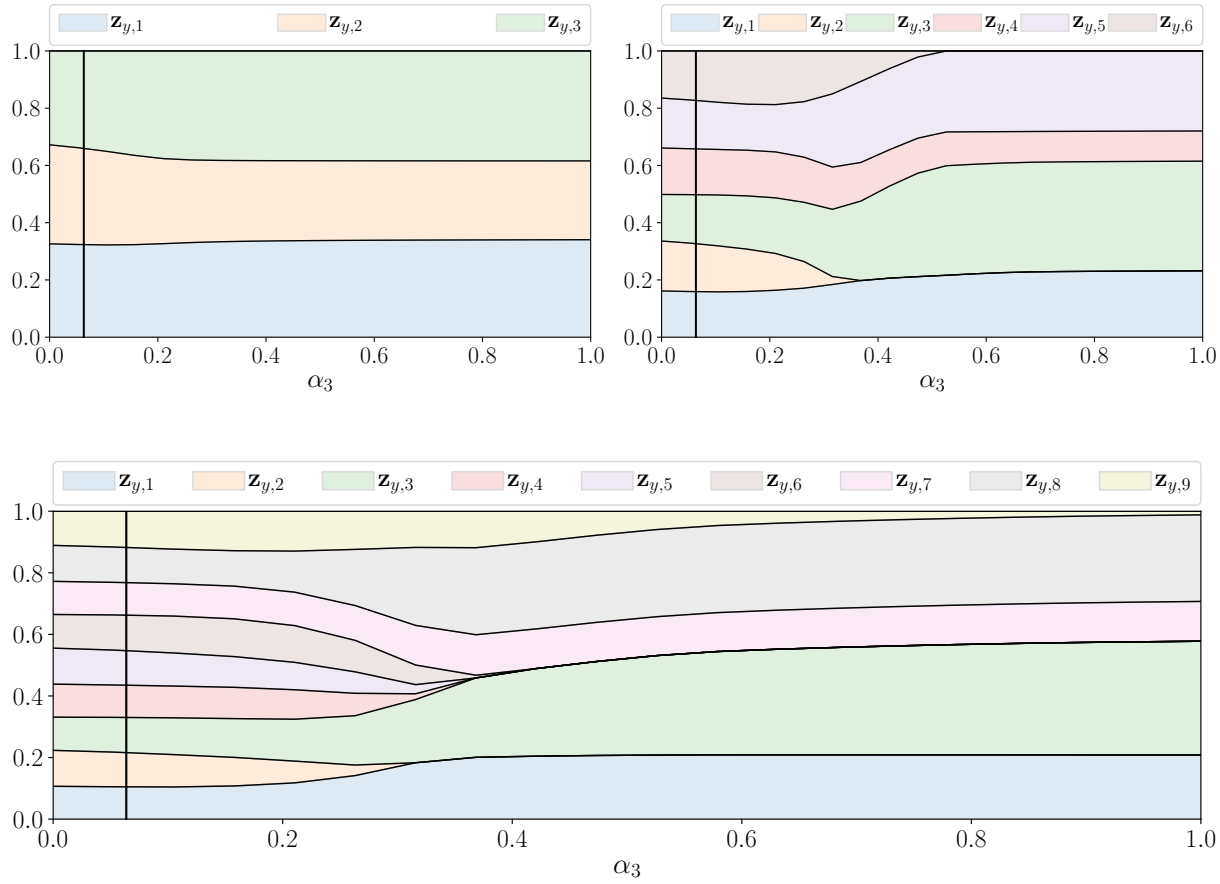


Figure 4: Target importances $\mathbf{z}_y$ with respect to α_3 for QPFS with Symmetric Importance

We compare the proposed strategies of the multivariate QPFS that are given in Table 1 for the ECoG dataset. First, we apply all the methods to obtain the feature importances. Then, we fit a linear model with an increasing number of features. For each method, the features are sorted by their obtained importances. We show how the described quality criteria change with the increasing feature set size. Figure 5 illustrates the results for the prediction of $k = 30$ timestamps. Here, the feature importance threshold τ is represented by colored ticks. These thresholds are larger for the proposed methods in comparison to the baseline RelAgg strategy. The SymImp strategy has the largest threshold, and it does not allow one to obtain a small feature subset. However, this strategy shows the best performance in terms of the sRMSE using the test data. The second performance is given

by AsymImp. All proposed algorithms give smaller test errors 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 the predictive quality and the size of selected feature subset.

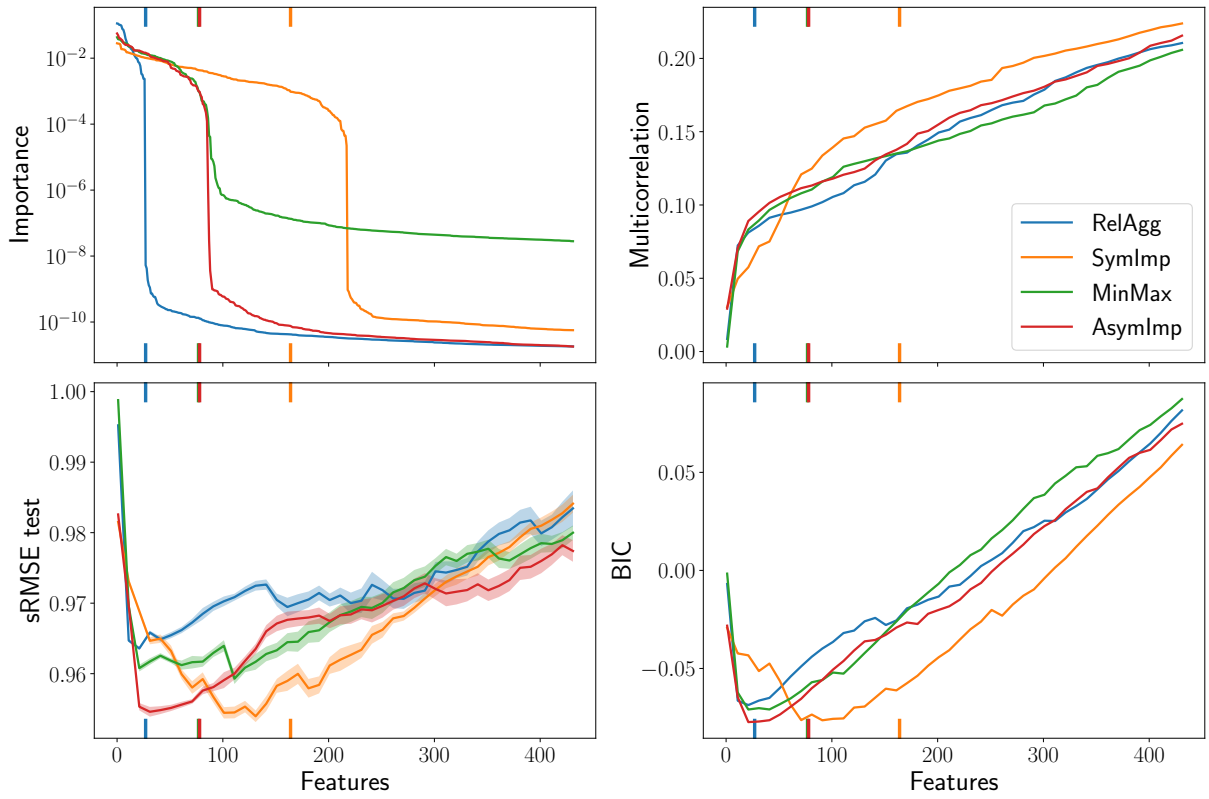


Figure 5: Feature selection algorithms evaluation for the ECoG data and the prediction of $k = 30$ timestamps

To compare the structure of the selected feature subsets and investigate the stability of the selection procedure, we use the bootstrap approach. First, the bootstrap data are generated. Then, we solve the feature selection problem for each pair of the design and target matrices. The obtained feature importances are compared. We calculate the average pairwise Spearman correlation coefficient and the ℓ_2 distance as the measures 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 model using the 50 features with the largest importances. AsymImp gives the least error on the test data. The size of the selected feature subsets is overestimated using the threshold $\tau = 10^{-4}$. The value of τ could be cross-validated to get the optimal threshold and feature subset size.

We fit the PLS regression model to the data to compare the dimensionality reduction and feature selection. Figure 6 demonstrates the scaled RMSE on the train and test data with respect to the dimensionality of latent space l. The test error reaches its minimum

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

at $l = 11$. The PLS regression is a more flexible approach compared to the linear model built on the subset of features. It results in the smallest error, but the model is not sparse.

Figure 7 compares 3 models: the linear regression and the PLS regression built on 100 features given the QPFS and the PLS regression with all features. We do not include the linear regression with all features because its results are close to the constant prediction. It also provides the result for the Lasso and Elastic Net algorithms that are widely used for feature selection. We use the AsymImp strategy for the QPFS in this experiment. The number of PLS latent dimension is $l = 15$. Here, the PLS regression is significantly better than the linear regression with the QPFS features. It means that the latter model is not flexible enough. However, the best result is by the PLS regression model combined with the 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 improves the model performance.

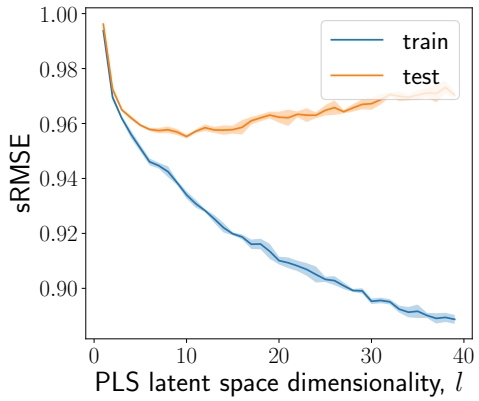


Figure 6: Test scaled RMSE for PLS regression models

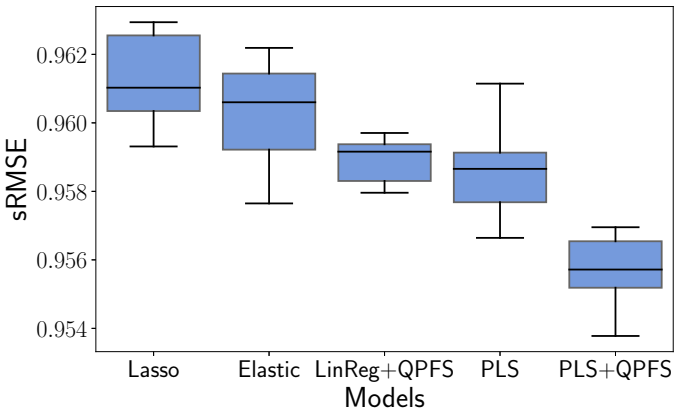


Figure 7: sRMSE box plots for different models

5. Conclusion

The study investigates the problem of signal decoding in which the data are highly redundant. To build a stable, adequate model, we reduced the dimensionality of the problem using the dependencies in both input and target spaces. The PLS regression is considered as a linear model for dimensionality reduction. The quadratic programming approach is

investigated as a 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 that are relevant to most difficult targets.

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

References

Agrawal, A., Verschueren, R., Diamond, S., Boyd, S., 2018. A rewriting system for convex optimization problems. Journal of Control and Decision 5, 42–60.

Biancolillo, A., Næs, T., Bro, R., Måge, I., 2017. Extension of SO-PLS to multi-way arrays: SO-N-PLS. Chemometrics and Intelligent Laboratory Systems 164, 113–126. doi:10.1016/j.chemolab.2017.03.002.

Blaschke, B., 1996. Some Newton type methods for the regularization of nonlinear ill-posed problems. Trauner.

Blaschke, B., Neubauer, A., Scherzer, O., 1997. On convergence rates for the iteratively regularized gauss-newton method. IMA Journal of Numerical Analysis 17, 421–436.

Chao, Z.C., Nagasaka, Y., Fujii, N., 2010. Long-term asynchronous decoding of arm motion using electrocorticographic signals in monkeys. Frontiers in Neuroengineering 3, 3. doi:10.3389/fneng.2010.00003.

Chiang, H.S., Shih, D.H., Lin, B., Shih, M.H., 2014. An apn model for arrhythmic beat classification. Bioinformatics 30, 1739–1746.

Chun, H., Keleş, S., 2010. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. Journal of the Royal Statistical Society. Series B: Statistical Methodology 72, 3–25. doi:10.1111/j.1467-9868.2009.00723.x.

Diamond, S., Boyd, S., 2016. CVXPY: A Python-embedded modeling language for convex optimization. Journal of Machine Learning Research 17, 1–5.

Ding, C., Peng, H., 2005. Minimum redundancy feature selection from microarray gene expression data. Journal of bioinformatics and computational biology 3, 185–205. doi:10.1142/s0219720005001004.

Eliseyev, A., Aksenova, T., 2014. Stable and artifact-resistant decoding of 3D hand trajectories from ECoG signals using the generalized additive model. Journal of Neural Engineering 11, 1–13. doi:10.1088/1741-2560/11/6/066005.

Eliseyev, A., Aksenova, T., 2016. Penalized multi-way partial least squares for smooth trajectory decoding from electrocorticographic (ECoG) recording. PLoS ONE 11, e0154878. doi:10.1371/journal.pone.0154878.

Eliseyev, A., Moro, C., Faber, J., Wyss, A., Torres, N., Mestais, C., Benabid, A.L., Aksenova, T., 2012. L1-Penalized N-way PLS for subset of electrodes selection in BCI experiments. Journal of Neural Engineering 9, 45010. doi:10.1088/1741-2560/9/4/045010.

Engel, S., Aksenova, T., Eliseyev, A., 2017. Kernel-based NPLS for continuous trajectory decoding from ECoG data for BCI applications, in: International Conference on Latent Variable Analysis and Signal Separation, Springer. pp. 417–426. doi:10.1007/978-3-319-53547-0_39.

Hervás, D., Prats-Montalbán, J.M., Lahoz, A., Ferrer, A., 2018. Sparse N-way partial least squares with R package sNPLS. Chemometrics and Intelligent Laboratory Systems 179, 54–63. doi:10.1016/j.chemolab.2018.06.005.

Isachenko, R., Vladimirova, M., Strijov, V., 2018. Dimensionality reduction for time series decoding and forecasting problems. DEStech Transactions on Computer Science and Engineering optim, 286–296.

Isachenko, R.V., Strijov, V.V., 2018. Quadratic programming optimization with feature selection for nonlinear models. Lobachevskii Journal of Mathematics 39, 1179–1187.

Karimi, S., Farrokhnia, M., 2014. 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. doi:10.1016/j.chemolab.2014.09.003.

Katrutsa, A., Strijov, V., 2015. Stress test procedure for feature selection algorithms. Chemometrics and Intelligent Laboratory Systems 142, 172–183. doi:10.1016/j.chemolab.2015.01.018.

Katrutsa, A., Strijov, V., 2017. Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria. Expert Systems with Applications 76, 1–11. doi:10.1016/j.eswa.2017.01.048.

Lauzon-Gauthier, J., Manolescu, P., Duchesne, C., 2018. The Sequential Multi-block PLS algorithm (SMB-PLS): Comparison of performance and interpretability. Chemometrics and Intelligent Laboratory Systems 180, 72–83. doi:10.1016/j.chemolab.2018.07.005.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J., Liu, H., 2017. Feature selection: A data perspective. ACM Computing Surveys 50, 94. doi:10.1145/3136625.

Li, X., Chen, L., Tang, Y., 2020. Hard: Bit-split string matching using a heuristic algorithm to reduce memory demand. Romanian Journal of Information Science and Technology 23, 94–105.

Lin, Y.W., Deng, B.C., Xu, Q.S., Yun, Y.H., Liang, Y.Z., 2016. The equivalence of partial least squares and principal component regression in the sufficient dimension reduction framework. Chemometrics and Intelligent Laboratory Systems 150, 58–64. doi:10.1016/j.chemolab.2015.11.003.

Mehmood, T., Sæbø, S., Liland, K.H., 2020. Comparison of variable selection methods in partial least squares regression. Journal of Chemometrics 118, 62–69. doi:10.1002/cem.3226.

Motrenko, A., Strijov, V., 2018. Multi-way Feature Selection for ECoG-based Brain-Computer Interface. Expert Systems with Applications 114, 402–413. doi:10.1016/j.eswa.2018.06.054.

Nesterov, Y., 1983. A method of solving a convex programming problem with convergence rate o (1/k2), in: Soviet Mathematics Doklady, pp. 372–376.

Prasad, Y., Biswas, K.K., Singla, P., 2013. Scaling-up quadratic programming feature selection, in: AAAI Workshop - Technical Report, pp. 95–97.

Precup, R.E., David, R.C., Roman, R.C., Szedlak-Stinean, A.I., Petriu, E.M., 2021. Optimal tuning of interval type-2 fuzzy controllers for nonlinear servo systems using slime mould algorithm. International Journal of Systems Science , 1–16.

Preitl, Z., Precup, R.E., Tar, J.K., Takács, M., 2006. Use of multi-parametric quadratic programming in fuzzy control systems. Acta Polytechnica Hungarica 3, 29–43.

Rodriguez-Lujan, I., Huerta, R., Elkan, C., Cruz, C.S., 2010. Quadratic programming feature selection. Journal of Machine Learning Research 11, 1491–1516.

Rosipal, R., 2010. Nonlinear partial least squares: An overview, in: Chemoinformatics and Advanced Machine Learning Perspectives: Complex Computational Methods and Collaborative Techniques. IGI Global, pp. 169–189. doi:10.4018/978-1-61520-911-8.ch009.

Rosipal, R., Krämer, N., 2006. Overview and recent advances in partial least squares. Lecture Notes in Computer Science 3940, 34–51. doi:10.1007/11752790_2.

Shimoda, K., Nagasaka, Y., Chao, Z.C., Fujii, N., 2012. Decoding continuous three-dimensional hand trajectories from epidural electrocorticographic signals in Japanese macaques. Journal of Neural Engineering 9, 36015. doi:10.1088/1741-2560/9/3/036015.

Yamada, M., Jitkrittum, W., Sigal, L., Xing, E.P., Sugiyama, M., 2014. High-Dimensional feature selection by feature-Wise kernelized lasso. Neural Computation 26, 185–207. doi:10.1162/NECO_a_00537.