1 Problem statement

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

$$\mathbf{y} = \mathbf{\Theta}\mathbf{x} + \boldsymbol{\varepsilon} \tag{1}$$

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

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

The columns $\boldsymbol{\chi}_j$ of the matrix $\mathbf{X}$ respond to object features. The examples of how to construct the dataset for a particular application task are described in Section Computational experiment.

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

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

The solution of the problem (2) is given by

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

The linear dependent columns of the matrix $\mathbf{X}$ leads to an instable solution for the optimization problem (2). If there is a vector $\boldsymbol{\alpha} \neq 0$ such that $\mathbf{X}\boldsymbol{\alpha} = 0$, then adding the vector $\boldsymbol{\alpha}$ to any column of the matrix $\mathbf{\Theta}$ does not change the error function $S(\mathbf{\Theta}|\mathbf{X}, \mathbf{Y})$. In this case the matrix $\mathbf{X}^T\mathbf{X}$ is not invertible. To avoid the strong linear dependence, feature selection and dimensionality reduction techniques are used.

2 Feature selection

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

$$\mathcal{A} = \underset{\mathcal{A}' \subseteq \{1, \dots, n\}}{\arg\max} \; Q(\mathcal{A}'|\mathbf{X}, \mathbf{Y}). \tag{3}$$

Once the solution $\mathcal{A}$ for the problem (3) is known, the problem (2) becomes

$$S(\mathbf{\Theta}_{\mathcal{A}}|\mathbf{X}_{\mathcal{A}}, \mathbf{Y}) = \left\| \mathbf{Y} - \mathbf{X}_{\mathcal{A}}\mathbf{\Theta}_{\mathcal{A}}^T \right\|_2^2 =\to \min_{\mathbf{\Theta}_{\mathcal{A}}}, \tag{4}$$

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

2.1 Quadratic Programming Feature Selection

One of the approach to the feature selection is to maximize feature relevances and minimize pairwise feature redundancy. The QPFS algorithm selects non-correlated features, which are relevant to the target vector $\boldsymbol{\phi}$ for the linear regression problem ($r = 1$)

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

Introduce two functions: $\text{Sim}(\mathbf{X})$ and $\text{Rel}(\mathbf{X}, \boldsymbol{\phi})$. The $\text{Sim}(\mathbf{X})$ measures the redundancy between features, the $\text{Rel}(\mathbf{X}, \boldsymbol{\phi})$ contains relevances between each feature and the target vector $\boldsymbol{\phi}$. We want to minimize the function Sim and maximize the Rel simultaneously.

QPFS offers the explicit way to construct the functions Sim and Rel. The method minimizes the following functional

$$(1 - \alpha) \cdot \underbrace{\mathbf{a}^T \mathbf{Q} \mathbf{a}}_{\text{Sim}} - \alpha \cdot \underbrace{\mathbf{b}^T \mathbf{a}}_{\text{Rel}} \to \min_{\substack{\mathbf{a} \in \mathbb{R}^n_+ \\ \|\mathbf{a}\|_1 = 1}}. \tag{5}$$

The matrix $\mathbf{Q} \in \mathbb{R}^{n \times n}$ entries measure the pairwise similarities between features. The vector $\mathbf{b} \in \mathbb{R}^n$ expresses the similarities between each feature and the target matrix $\mathbf{b}$. The normalized vector $\mathbf{a}$ shows the importance of each feature. The functional (5) penalizes the dependent features by the function Sim and encourages features relevant to the target by the function Rel. The parameter α allows to control the trade-off between the functions Sim and the Rel. The authors of the original QPFS paper suggested the way to select α and make $\text{Sim}(\mathbf{X})$ and $\text{Rel}(\mathbf{X}, \boldsymbol{\phi})$ impact the same

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

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

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

To measure similarity the authors use the absolute value of sample correlation coefficient between pairs of features for the function Sim, and between features and the target vector $\boldsymbol{\phi}$ for the function Rel

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

The problem (5) is convex if the matrix $\mathbf{Q}$ is positive semidefinite. In general it is not always true. To satisfy this condition, the matrix $\mathbf{Q}$ spectrum is shifted and the matrix $\mathbf{Q}$ is replaced by $\mathbf{Q} - \lambda_{\min}\mathbf{I}$, where $\lambda_{\min}$ is a $\mathbf{Q}$ minimal eigenvalue.

The functional (5) corresponds to the quality criteria $Q(\mathcal{A}|\mathbf{X}, \boldsymbol{\phi})$

$$\mathcal{A} = \mathop{\arg\max}_{\mathcal{A}' \subseteq \{1,\dots,n\}} Q(\mathcal{A}'|\mathbf{X}, \boldsymbol{\phi}) \Leftrightarrow \mathop{\arg\min}_{\mathbf{a} \in \mathbb{R}^n_+, \, \|\mathbf{a}\|_1 = 1} \left[\mathbf{a}^T \mathbf{Q} \mathbf{a} - \alpha \cdot \mathbf{b}^T \mathbf{a} \right]. \tag{7}$$

2.2 Multivariate QPFS

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

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

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

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

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

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

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

To take into account the dependencies in the columns of the matrix $\mathbf{Y}$ we extend the QPFS functional (5) to the multivariate case. We add the term $\text{Sim}(\mathbf{Y})$ and extend the term $\text{Rel}(\mathbf{X}, \mathbf{Y})$:

$$\alpha_1 \cdot \underbrace{\mathbf{a}_x^T \mathbf{Q}_x \mathbf{a}_x}_{\text{Sim}(\mathbf{X})} - \alpha_2 \cdot \underbrace{\mathbf{a}_x^T \mathbf{B} \mathbf{a}_y}_{\text{Rel}(\mathbf{X},\mathbf{Y})} + \alpha_3 \cdot \underbrace{\mathbf{a}_y^T \mathbf{Q}_y \mathbf{a}_y}_{\text{Sim}(\mathbf{Y})} \rightarrow \min_{\substack{\mathbf{a}_x \in \mathbb{R}^n_+ \, \|\mathbf{a}_x\|_1 = 1 \\ \mathbf{a}_y \in \mathbb{R}^r_+ \, \|\mathbf{a}_y\|_1 = 1}}. \tag{8}$$

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

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

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

Statement 1. *For the case $r = 1$ the proposed functional* (8) *coincides with the original QPFS algorithm* (5)*.*

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

$$\alpha_1 \cdot \mathbf{a}_x^T \mathbf{Q}_x \mathbf{a}_x - \alpha_2 \cdot \mathbf{a}_x^T \mathbf{b} \to \min_{\mathbf{a}_x \in \mathbb{R}^n_+ \, \|\mathbf{a}_x\|_1 = 1}.$$

Setting $\alpha = \frac{\alpha_2}{\alpha_1 + \alpha_2}$ brings to the original QPFS problem (5). $\square$

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

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

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.

Statement 2. *Balance between the terms Sim*$(\mathbf{X})$ *and Rel*$(\mathbf{X}, \mathbf{Y})$ *for the problem* (8) *is achieved by the following coefficients:*

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

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

Proof. The impact of these terms are equal if $\alpha_1 \cdot \text{Sim}(\mathbf{X}) = \alpha_2 \cdot \text{Rel}(\mathbf{X}, \mathbf{Y})$. The mean values of the terms $\text{Sim}(\mathbf{X})$ and $\text{Rel}(\mathbf{X}, \mathbf{Y})$ are given by the mean values $\overline{\mathbf{Q}}_x$ and $\overline{\mathbf{B}}$ of the corresponding matrices $\mathbf{Q}_x$ and $\mathbf{B}$. Since $\alpha_1 + \alpha_2 + \alpha_3 = 1$, we obtain $(1 - \alpha_3 - \alpha_2)\overline{\mathbf{Q}}_x = \alpha_2 \overline{\mathbf{B}}$. Express α_2 to get

$$\alpha_2 = \frac{(1 - \alpha_3)\overline{\mathbf{Q}}_x}{\overline{\mathbf{Q}}_x + \overline{\mathbf{B}}}.$$

The value for α_1 is derived from the $\alpha_1 + \alpha_2 + \alpha_3 = 1$. $\square$

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

3 Feature categorization

Feature selection algorithms eliminate features which are not relevant to the target variable. To determine whether the feature is relevant the t-test could be applied for the correlation coefficient.

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

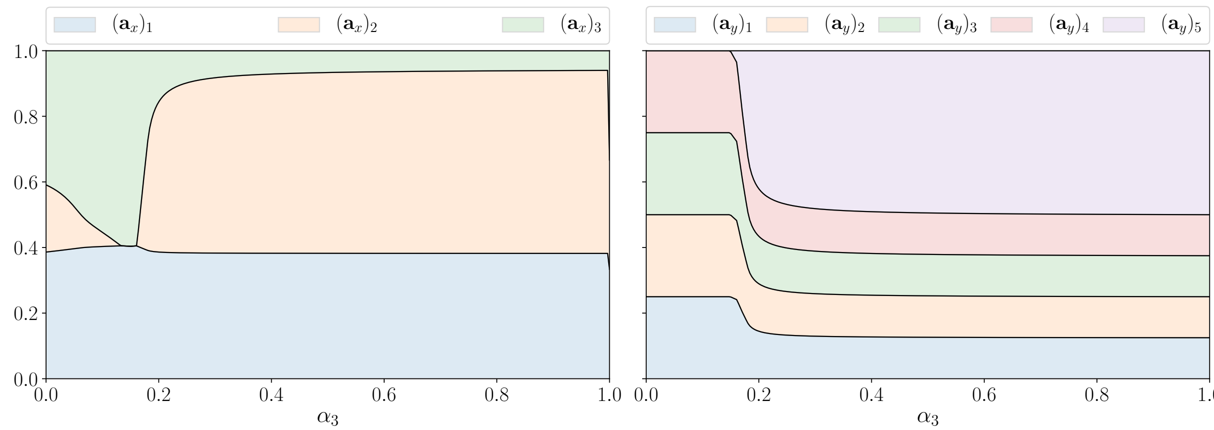

Figure 1: Feature importances $\mathbf{a}_x$ and $\mathbf{a}_y$ with respect to the α_3 coefficient

$$H_0 : r = 0$$
$$H_1 : r \neq 0$$

If features are relevant, but correlated, feature selection methods pick the subset of them to reduce the multicollinearity and redundancy. The goal is to find relevant, non-correlated features. However, in this case the correlations between targets in matrix $\mathbf{Y}$ are crucial. To measure the dependence of each feature or target, the Variance Inflation Factor is computed

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

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

On that basis, we categorize features into 5 disjoint groups:

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

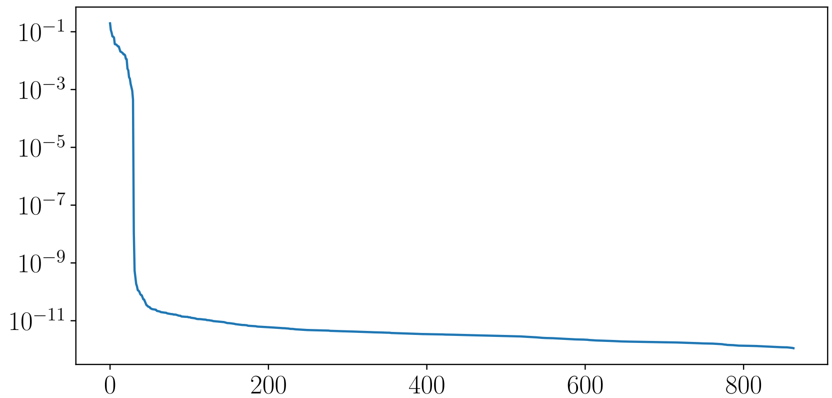

Figure 2: Sorted feature importances for the QPFS algorithm

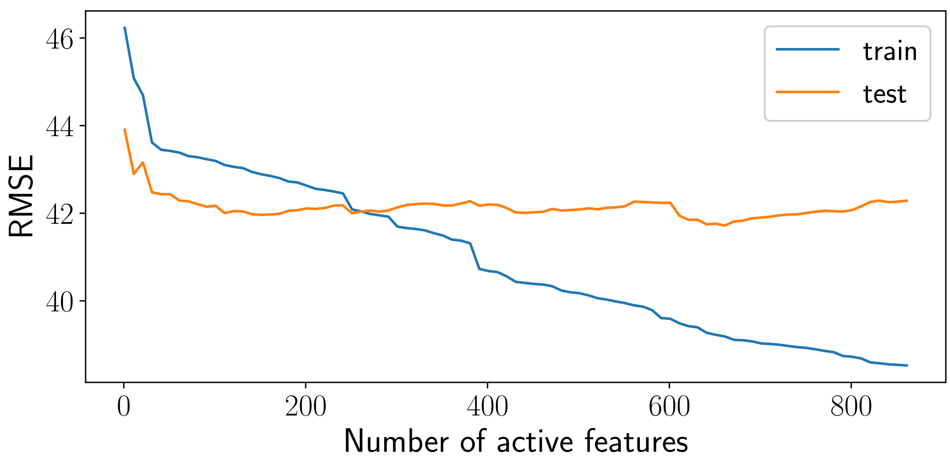

Figure 3: RMSE with respect to size of active set, features are ranked by QPFS algorithm

4 Experiment

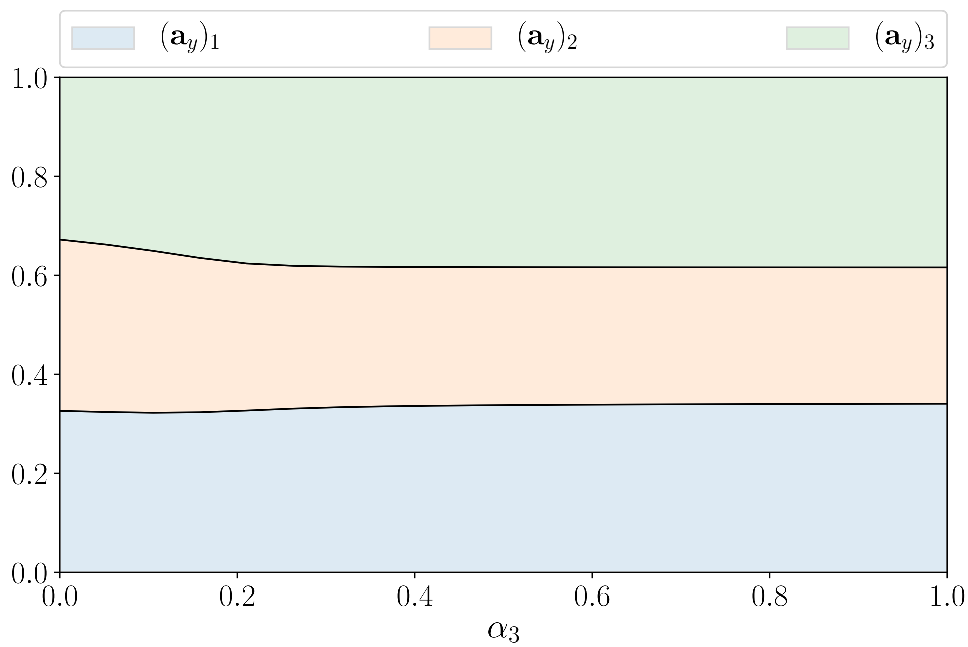

Figure 4: Targets importances for ECoG data, each component is related to one of the axis