# 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

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

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

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

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:

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

The solution of the problem (2) is given by

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

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 $\boldsymbol{\Theta}$ does not change the error function $S(\boldsymbol{\Theta}|\mathbf{X}, \mathbf{Y})$. 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 $\mathcal{A} = \{1, \ldots, n\}$ of the matrix $\mathbf{X}$ columns. To select the set $\mathcal{A}$ among all possible $2^n - 1$ subsets, introduce the feature selection quality criteria

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

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

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

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 $\boldsymbol{\nu}$ for the linear regression problem ($r = 1$)

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

Introduce two functions: $\mathrm{Sim}(\mathbf{X})$ and $\mathrm{Rel}(\mathbf{X}, \boldsymbol{\nu})$. The $\mathrm{Sim}(\mathbf{X})$ measures the redundancy between features, the $\mathrm{Rel}(\mathbf{X}, \boldsymbol{\nu})$ contains relevances between each feature and the target vector $\boldsymbol{\nu}$. 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

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

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 $\alpha$ 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 $\alpha$ and make $\mathrm{Sim}(\mathbf{X})$ and $\mathrm{Rel}(\mathbf{X}, \boldsymbol{\nu})$ impact the same

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

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

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

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 $\boldsymbol{\nu}$ for the function Rel

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

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 $Q(\mathcal{A}|\mathbf{X}, \boldsymbol{\nu})$

$$\mathcal{A} = \arg\max_{\mathcal{A}'\subseteq\{1,\ldots,n\}} Q(\mathcal{A}'|\mathbf{X}, \boldsymbol{\nu}) \Leftrightarrow \arg\min_{\mathbf{a}\in\mathbb{R}^n_+,\, \mathbf{1}_n^T\mathbf{a}=1} \left[\mathbf{a}^T\mathbf{Q}\mathbf{a} - \alpha\cdot\mathbf{b}^T\mathbf{a}\right]. \tag{7}$$

## 2.2 Multivariate QPFS

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

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

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

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

We have three features and $r$ targets, where first $r-1$ target are the identical. The pairwise features similarities are given by the matrix $\mathbf{Q}$. 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.

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

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

**Noname** 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 $\mathrm{Sim}(\mathbf{Y})$ and extend the term $\mathrm{Rel}(\mathbf{X}, \mathbf{Y})$:

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

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

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

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

**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

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

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

The coefficients $\alpha_1$, $\alpha_2$, and $\alpha_3$ 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 $\mathrm{Sim}(\mathbf{X})$ and $\mathrm{Rel}(\mathbf{X}, \mathbf{Y})$ by fixing the proportion between $\alpha_1$ and $\alpha_2$.

**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 $\alpha_1 \cdot \mathrm{Sim}(\mathbf{X}) = \alpha_2 \cdot \mathrm{Rel}(\mathbf{X}, \mathbf{Y})$. The mean values of the terms $\mathrm{Sim}(\mathbf{X})$ and $\mathrm{Rel}(\mathbf{X}, \mathbf{Y})$ 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 $\alpha_2$ to get

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

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

**Statement 3.** *Balance between the terms Sim*$(\mathbf{X})$*, Rel*$(\mathbf{X}, \mathbf{Y})$*, and Rel*$(\mathbf{X}, \mathbf{Y})$ *for the problem* (8) *is achieved by the following coefficients:*

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

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

*Proof.* The mean values of the terms $\mathrm{Sim}(\mathbf{X})$, $\mathrm{Rel}(\mathbf{X}, \mathbf{Y})$, and $\mathrm{Rel}(\mathbf{X}, \mathbf{Y})$ are given by the mean values $\overline{\mathbf{Q}}_x$, $\overline{\mathbf{B}}$, $\overline{\mathbf{Q}}_y$ of the corresponding matrices $\mathbf{Q}_x$, $\mathbf{B}$, and $\mathbf{Q}_y$. The desired values of $\alpha_1$, $\alpha_2$, and $\alpha_3$ are given by solution of the following equations

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

□

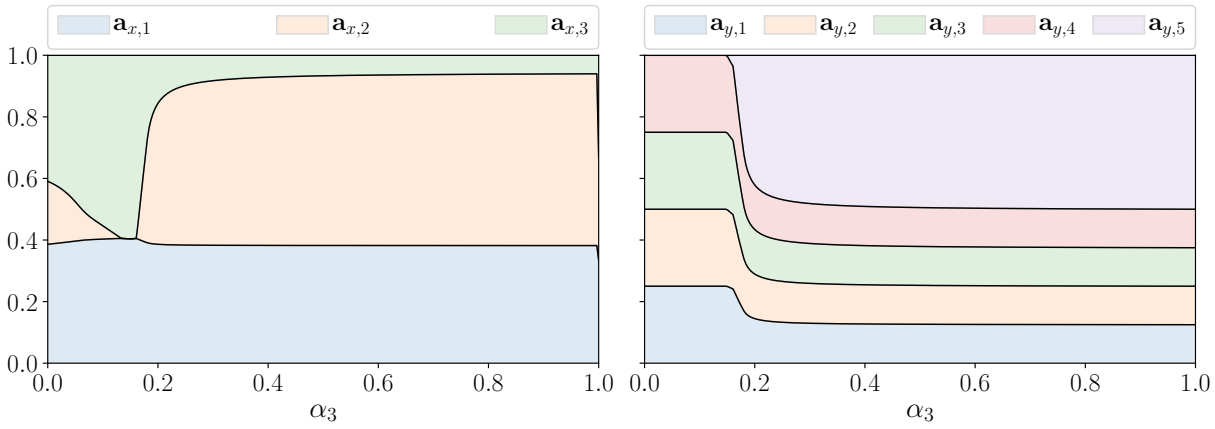



Figure 1: Feature importances $\mathbf{a}_x$ and $\mathbf{a}_y$ with respect to the $\alpha_3$ coefficient

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 $\mathrm{corr}(\boldsymbol{\nu}_1, \boldsymbol{\nu}_2) = 0.2$ and all others entries to one. Figure 1 shows the importances of features $\mathbf{a}_x$ and targets $\mathbf{a}_y$ with respect to $\alpha_3$ coefficient. If $\alpha_3$ is small, the impact of all targets are almost equal and the feature $\boldsymbol{\chi}_3$ dominates the feature $\boldsymbol{\chi}_2$. When $\alpha_3$ becomes larger than 0.2, the importance $(\mathbf{a}_y)_5$ of the target $\phi_5$ grows up along with the importance of the feature $\boldsymbol{\chi}_2$.

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

$$\begin{aligned}
\alpha_1 \cdot \underbrace{\mathbf{a}_x^T \mathbf{Q}_x \mathbf{a}_x}_{\mathrm{Sim}(\mathbf{X})} - \alpha_2 \cdot \underbrace{\mathbf{a}_x^T \mathbf{B} \mathbf{a}_y}_{\mathrm{Rel}(\mathbf{X},\mathbf{Y})} &\to \min_{\substack{\mathbf{a}_x \in \mathbb{R}^n_+ \\ \mathbf{1}_n^T \mathbf{a}_x = 1}}; \\
\alpha_3 \cdot \underbrace{\mathbf{a}_y^T \mathbf{Q}_y \mathbf{a}_y}_{\mathrm{Sim}(\mathbf{Y})} + \alpha_2 \cdot \underbrace{\mathbf{a}_x^T \mathbf{B} \mathbf{a}_y}_{\mathrm{Rel}(\mathbf{X},\mathbf{Y})} &\to \min_{\substack{\mathbf{a}_y \in \mathbb{R}^r_+ \\ \mathbf{1}_r^T \mathbf{a}_y = 1}}.
\end{aligned}$$

It brings us to the min-max or max-min formulation

$$\max_{\substack{\mathbf{a}_y\in\mathbb{R}^r_+\\ \mathbf{1}_r^T\mathbf{a}_y=1}}\min_{\substack{\mathbf{a}_x\in\mathbb{R}^n_+\\ \mathbf{1}_n^T\mathbf{a}_x=1}}\left[\alpha_1\cdot\underbrace{\mathbf{a}_x^T\mathbf{Q}_x\mathbf{a}_x}_{\mathrm{Sim}(\mathbf{X})}-\alpha_2\cdot\underbrace{\mathbf{a}_x^T\mathbf{B}\mathbf{a}_y}_{\mathrm{Rel}(\mathbf{X},\mathbf{Y})}-\alpha_3\cdot\underbrace{\mathbf{a}_y^T\mathbf{Q}_y\mathbf{a}_y}_{\mathrm{Sim}(\mathbf{Y})}\right]; \tag{9}$$

$$\min_{\substack{\mathbf{a}_x\in\mathbb{R}^n_+\\ \mathbf{1}_n^T\mathbf{a}_x=1}}\max_{\substack{\mathbf{a}_y\in\mathbb{R}^r_+\\ \mathbf{1}_r^T\mathbf{a}_y=1}}\left[\alpha_1\cdot\underbrace{\mathbf{a}_x^T\mathbf{Q}_x\mathbf{a}_x}_{\mathrm{Sim}(\mathbf{X})}-\alpha_2\cdot\underbrace{\mathbf{a}_x^T\mathbf{B}\mathbf{a}_y}_{\mathrm{Rel}(\mathbf{X},\mathbf{Y})}-\alpha_3\cdot\underbrace{\mathbf{a}_y^T\mathbf{Q}_y\mathbf{a}_y}_{\mathrm{Sim}(\mathbf{Y})}\right]. \tag{10}$$

**Theorem 1.** *For positive definite matrices* $\mathbf{Q}_x$ *and* $\mathbf{Q}_y$ *the problems* (9) *and* (10) *have the same optimal value.*

*Proof.* Denote

$$\mathbb{C}^n=\{\mathbf{a}:\mathbf{a}\in\mathbb{R}^n_+,\ \mathbf{1}_n^T\mathbf{a}=1\},\ \mathbb{C}^r=\{\mathbf{a}:\mathbf{a}\in\mathbb{R}^r_+,\ \mathbf{1}_r^T\mathbf{a}=1\};$$
$$f(\mathbf{a}_x,\mathbf{a}_y)=\alpha_1\cdot\mathbf{a}_x^T\mathbf{Q}_x\mathbf{a}_x-\alpha_2\cdot\mathbf{a}_x^T\mathbf{B}\mathbf{a}_y-\alpha_3\cdot\mathbf{a}_y^T\mathbf{Q}_y\mathbf{a}_y.$$

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

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

□

Let solve the problem (9). Fix some $\mathbf{a}_y\in\mathbb{C}^r$. The Lagrangian of this problem is

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

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

$$\max_{\lambda,\,\boldsymbol{\mu}\in\mathbb{R}^n_+}g(\mathbf{a}_y,\lambda,\boldsymbol{\mu})=\max_{\lambda,\,\boldsymbol{\mu}\in\mathbb{R}^n_+}\left[\min_{\mathbf{a}_x\in\mathbb{R}^n}L(\mathbf{a}_x,\mathbf{a}_y,\lambda,\boldsymbol{\mu})\right]$$

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

$$\mathbf{a}_x=\frac{1}{2\alpha_1}\mathbf{Q}_x^{-1}\left(\alpha_2\cdot\mathbf{B}\mathbf{a}_y+\lambda\cdot\mathbf{1}+\boldsymbol{\mu}\right).$$

The dual function is equal to

$$g(\mathbf{a}_y,\lambda,\boldsymbol{\mu})=\min_{\mathbf{a}_x\in\mathbb{R}^n}L(\mathbf{a}_x,\mathbf{a}_y,\lambda,\boldsymbol{\mu})=\frac{1}{4\alpha_1^2}\left(\alpha_2\cdot\mathbf{B}\mathbf{a}_y+\lambda\cdot\mathbf{1}+\boldsymbol{\mu}\right)^T\mathbf{Q}_x^{-1}\left(\alpha_2\cdot\mathbf{B}\mathbf{a}_y+\lambda\cdot\mathbf{1}+\boldsymbol{\mu}\right)$$
$$-\frac{1}{2\alpha_1}\mathbf{Q}_x^{-1}\left(\alpha_2\cdot\mathbf{B}\mathbf{a}_y+\lambda\cdot\mathbf{1}+\boldsymbol{\mu}\right)^T\mathbf{Q}_x^{-1}\mathbf{B}\mathbf{a}_y-\lambda\left(\frac{1}{2\alpha_1}\mathbf{1}_n^T\mathbf{Q}_x^{-1}\left(\alpha_2\cdot\mathbf{B}\mathbf{a}_y+\lambda\cdot\mathbf{1}+\boldsymbol{\mu}\right)-1\right)$$
$$-\frac{1}{2\alpha_1}\boldsymbol{\mu}^T\mathbf{Q}_x^{-1}\left(\alpha_2\cdot\mathbf{B}\mathbf{a}_y+\lambda\cdot\mathbf{1}+\boldsymbol{\mu}\right)=$$

# 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.

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

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

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

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

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

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

1. non-relevant features

$$\left\{j : \mathrm{corr}(\boldsymbol{\chi}_j, \boldsymbol{\nu}_k) = 0,\ \forall k \in \{1, \ldots, r\}\right\};$$

2. non-$\mathbf{X}$-correlated features, which are relevant to non-$\mathbf{Y}$-correlated targets

$$\left\{j : \big(\mathrm{VIF}(\boldsymbol{\chi}_j) < 10\big) \text{ and } \big(\mathrm{VIF}(\boldsymbol{\nu}_k) < 10,\ \forall k \in \{1, \ldots, r\} : \mathrm{corr}(\boldsymbol{\chi}_j, \boldsymbol{\nu}_k) \neq 0\big)\right\};$$

3. non-$\mathbf{X}$-correlated features, which are relevant to $\mathbf{Y}$-correlated targets

$$\left\{j : \big(\mathrm{VIF}(\boldsymbol{\chi}_j) < 10\big) \text{ and } \big(\exists k \in \{1, \ldots, r\} : \mathrm{VIF}(\boldsymbol{\nu}_k) > 10 \ \&\ \mathrm{corr}(\boldsymbol{\chi}_j, \boldsymbol{\nu}_k) \neq 0\big)\right\};$$

4. $\mathbf{X}$-correlated features, which are relevant to non-$\mathbf{Y}$-correlated targets

$$\left\{j : \big(\mathrm{VIF}(\boldsymbol{\chi}_j) > 10\big) \text{ and } \big(\mathrm{VIF}(\boldsymbol{\nu}_k) < 10,\ \forall k \in \{1, \ldots, r\} : \mathrm{corr}(\boldsymbol{\chi}_j, \boldsymbol{\nu}_k) \neq 0\big)\right\};$$

5. $\mathbf{X}$-correlated features, which are relevant to $\mathbf{Y}$-correlated targets

$$\left\{j : \big(\mathrm{VIF}(\boldsymbol{\chi}_j) > 10\big) \text{ and } \big(\exists k \in \{1, \ldots, r\} : \mathrm{VIF}(\boldsymbol{\nu}_k) > 10 \ \&\ \mathrm{corr}(\boldsymbol{\chi}_j, \boldsymbol{\nu}_k) \neq 0\big)\right\}.$$

# 4 Experiment

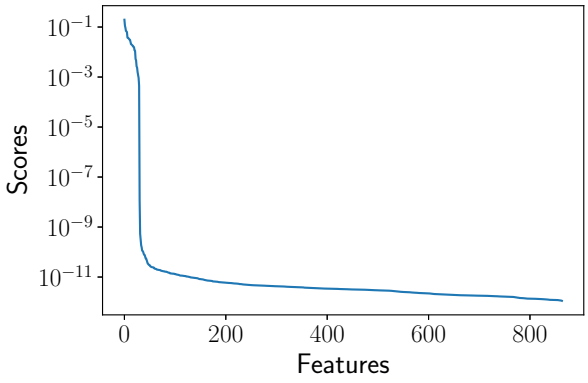



Figure 2: Sorted feature importances for the QPFS algorithm

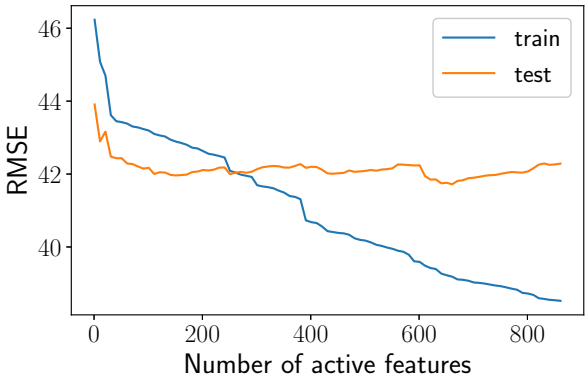



Figure 3: RMSE w.r.t. size of active set, features are ranked by QPFS algorithm

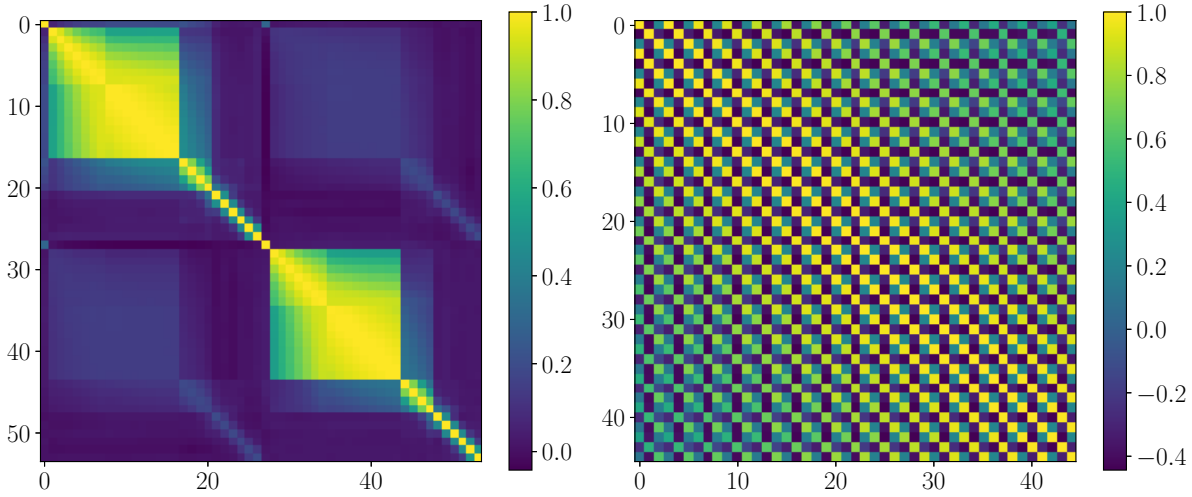
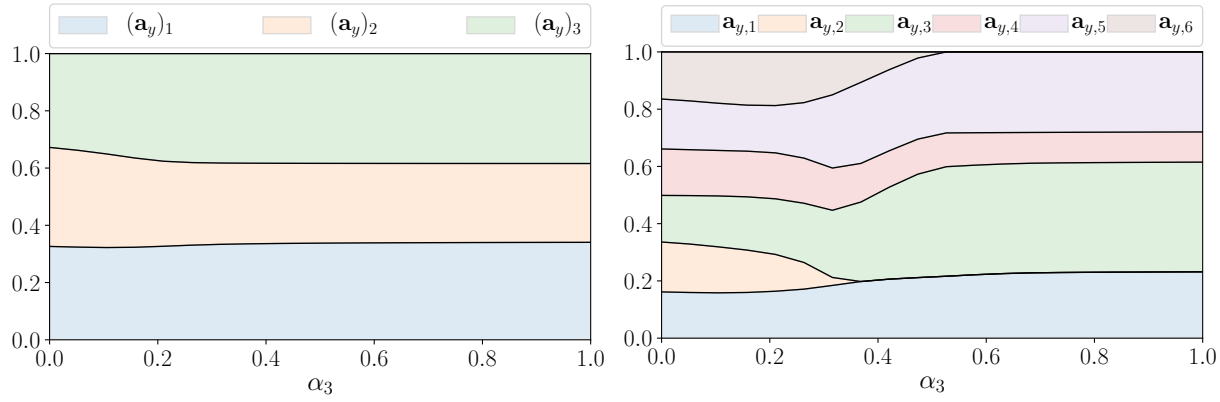


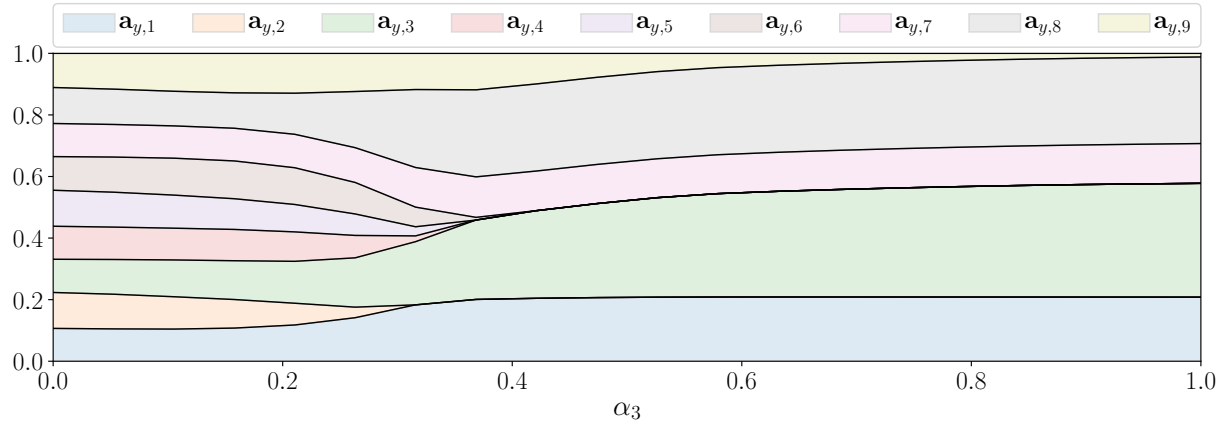

Figure 4: Correlation matrices for $\mathbf{X}$ and $\mathbf{Y}$




(a) autoregression step=1




(b) autoregression step=2




(c) autoregression step=3

Figure 5: Target importances $\mathbf{a}_y$ for ECoG data with respect to the $\alpha_3$ coefficient

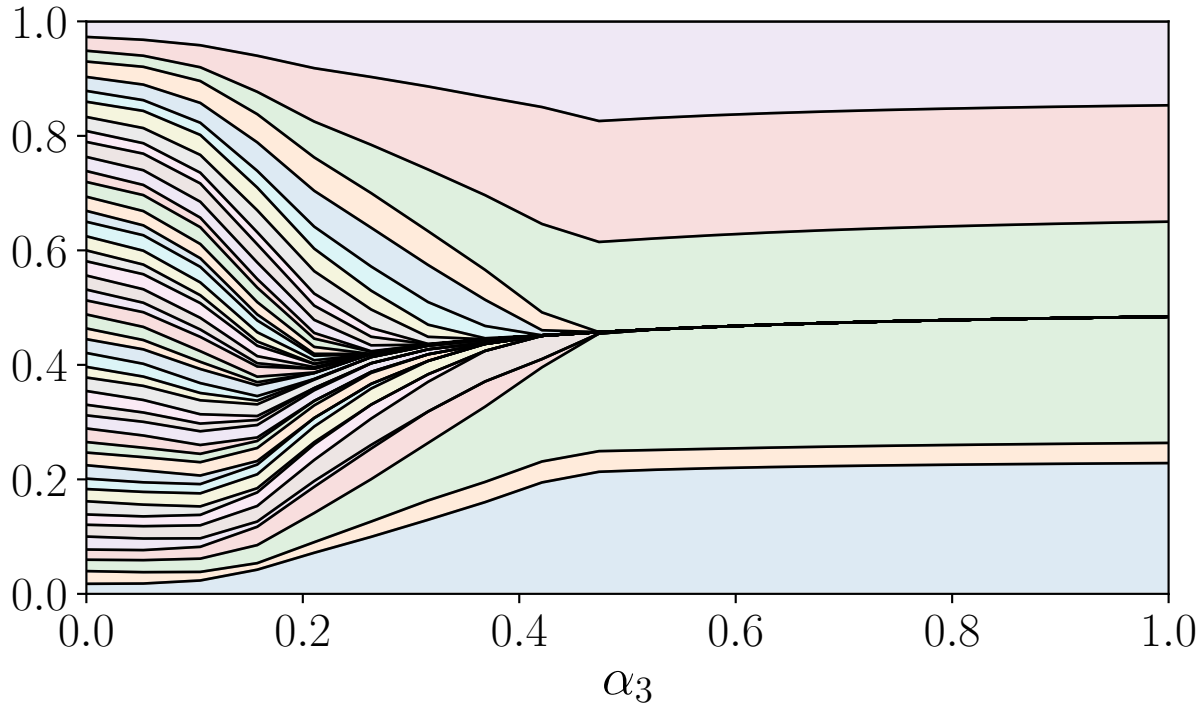



Figure 6: autoregression step=45

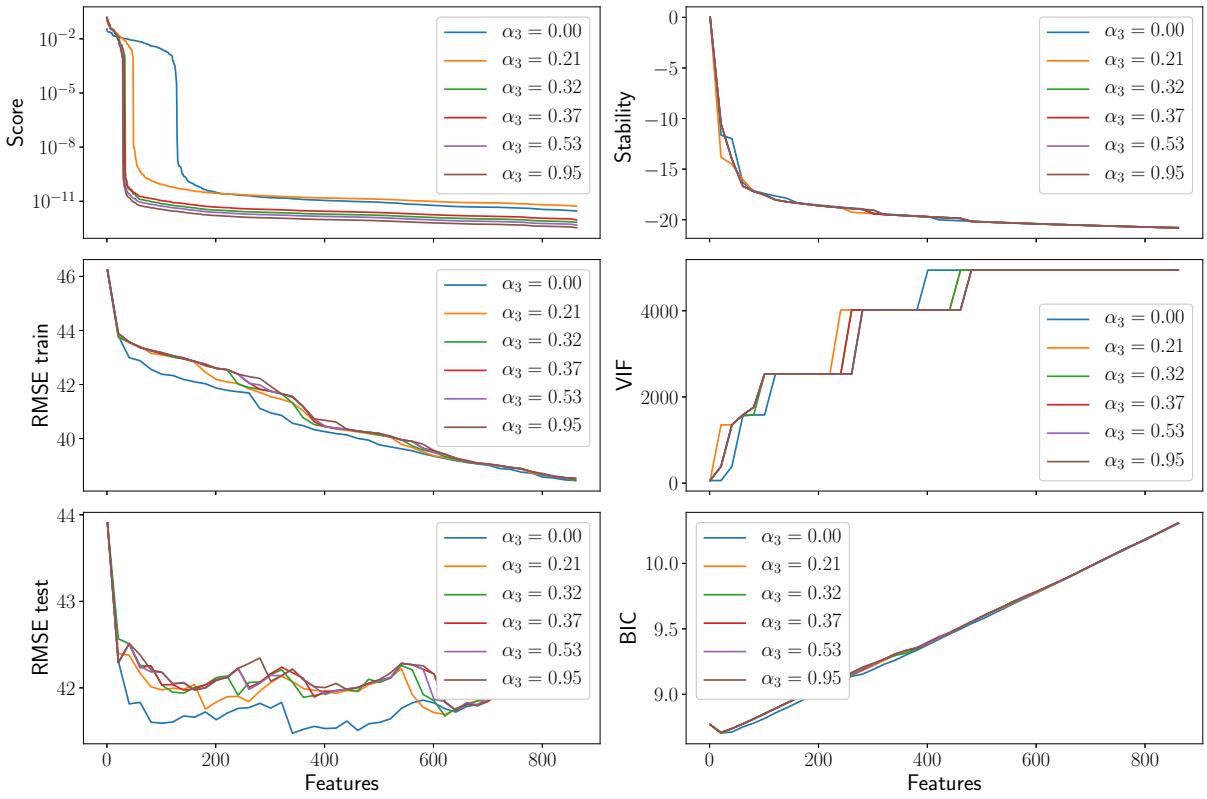

Score
Stability
RMSE train
VIF
RMSE test
BIC
Features
Features
$\alpha_3 = 0.00$
$\alpha_3 = 0.21$
$\alpha_3 = 0.32$
$\alpha_3 = 0.37$
$\alpha_3 = 0.53$
$\alpha_3 = 0.95$

Figure 7

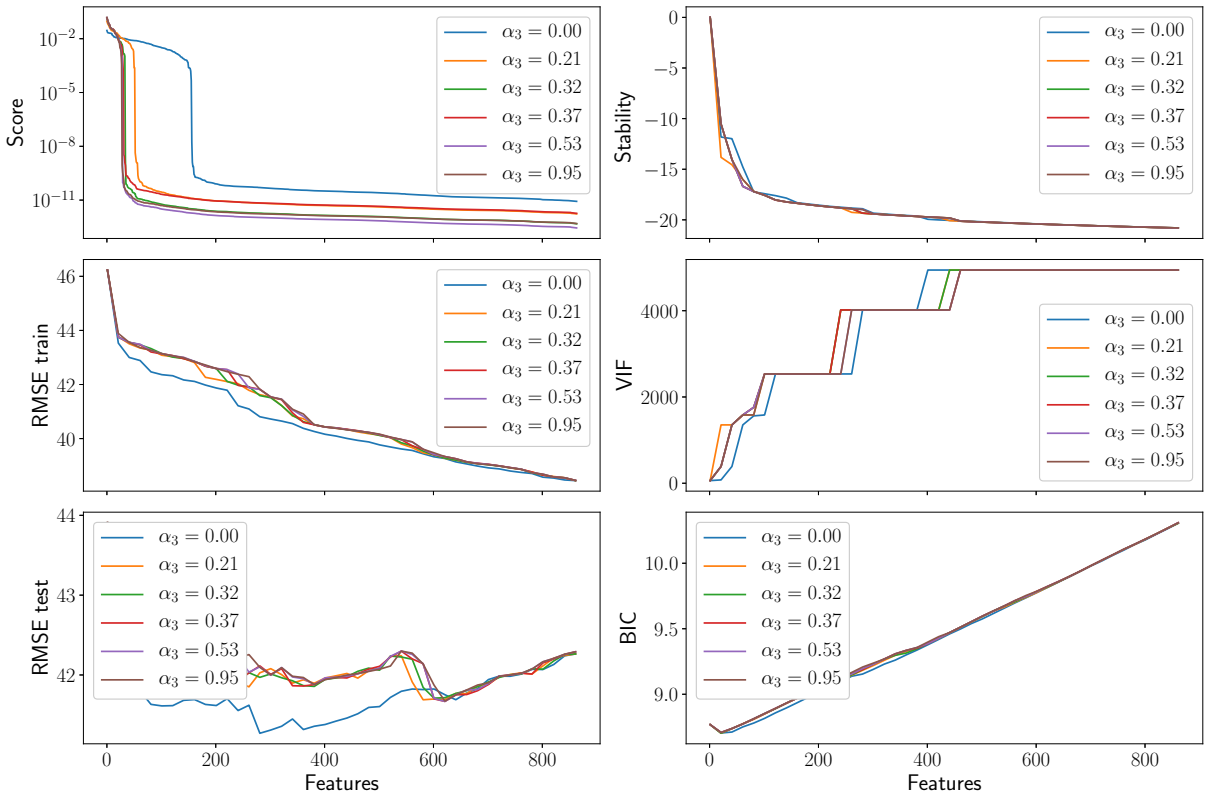

Score
Stability
RMSE train
VIF
RMSE test
BIC
Features
Features
$\alpha_3 = 0.00$
$\alpha_3 = 0.21$
$\alpha_3 = 0.32$
$\alpha_3 = 0.37$
$\alpha_3 = 0.53$
$\alpha_3 = 0.95$

Figure 8

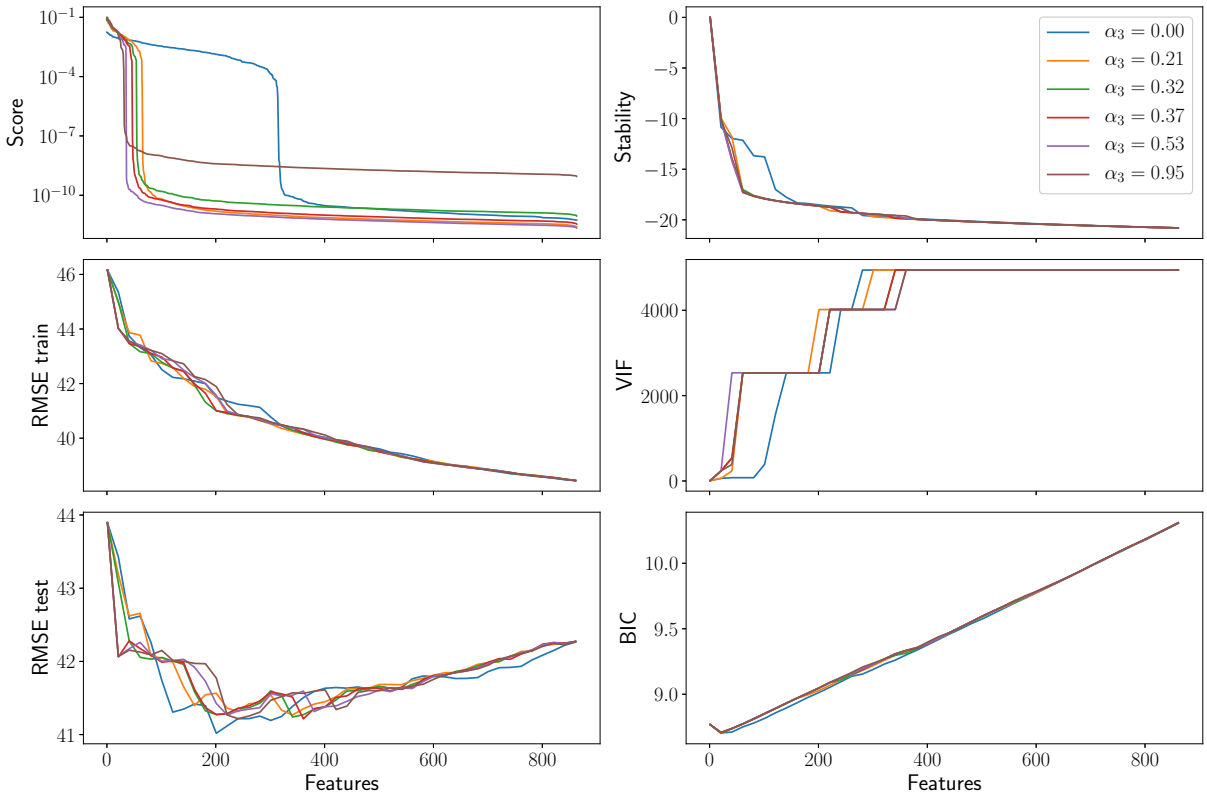

Score
Stability
RMSE train
VIF
RMSE test
BIC
Features
Features
$\alpha_3 = 0.00$
$\alpha_3 = 0.21$
$\alpha_3 = 0.32$
$\alpha_3 = 0.37$
$\alpha_3 = 0.53$
$\alpha_3 = 0.95$

Figure 9

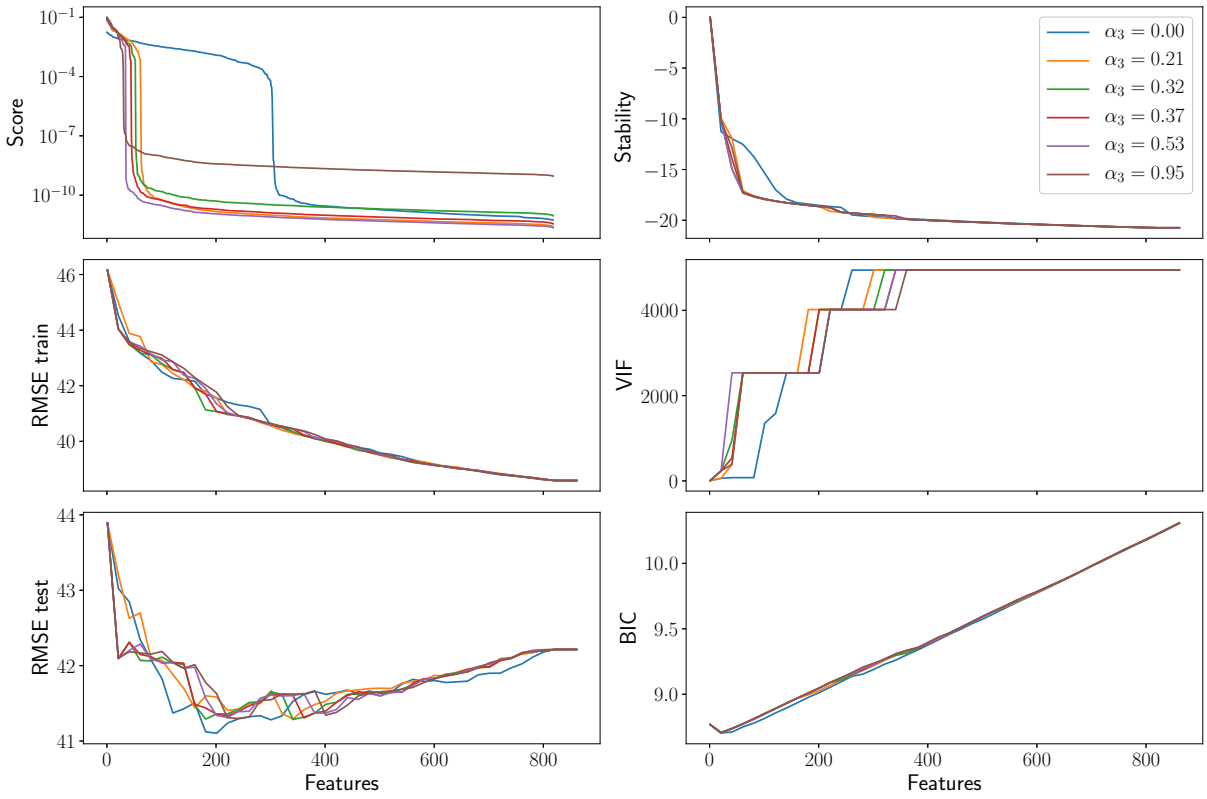

Score
Stability
RMSE train
VIF
RMSE test
BIC
Features
Features
$\alpha_3 = 0.00$
$\alpha_3 = 0.21$
$\alpha_3 = 0.32$
$\alpha_3 = 0.37$
$\alpha_3 = 0.53$
$\alpha_3 = 0.95$

Figure 10