1 Problem statement

The goal is to forecast a dependent target variable $\mathbf{y} \in \mathbb{R}^r$ from a independent input object $\mathbf{x} \in \mathbb{R}^n$. We assume that there is a linear dependence between the objects $\mathbf{x}$ and the target vector $\mathbf{y}$

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

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

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

The examples of how to construct the dataset for a particular application task are described in the Computational experiment.

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

$$S(\mathbf{\Theta}|\mathbf{X}, \mathbf{Y}) = \left\| \underset{m \times n}{\mathbf{X}} \cdot \underset{n \times r}{\mathbf{\Theta}} - \underset{m \times r}{\mathbf{Y}} \right\|_2^2 = \sum_{i=1}^{m} \left\| \underset{1 \times n}{\mathbf{x}_i} \cdot \underset{n \times r}{\mathbf{\Theta}} - \underset{1 \times r}{\mathbf{y}_i} \right\|_2^2 \to \min_{\mathbf{\Theta}}. \tag{2}$$

The solution of the problem (2) is given by

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

The linear dependence of the matrix $\mathbf{X}$ columns leads to an instable solution for the optimization problem (2). If there is a vector $\boldsymbol{\alpha} \neq 0$ such that $\mathbf{X}\boldsymbol{\alpha} = 0$ than 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 $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ is not invertible. To avoid the strong linear dependence feature selection and dimensionality reduction techniques are used.

2 Feature selection

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

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

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: $\mathrm{Sim}(\mathbf{X})$ and $\mathrm{Rel}(\mathbf{X}, \boldsymbol{\phi})$. The $\mathrm{Sim}(\mathbf{X})$ measures the redundancy between features, the $\mathrm{Rel}(\mathbf{X}, \boldsymbol{\phi})$ 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

$$(1-\alpha) \cdot \underbrace{\mathbf{a}^{\mathsf{T}}\mathbf{Q}\mathbf{a}}_{\mathrm{Sim}} - \alpha \cdot \underbrace{\mathbf{b}^{\mathsf{T}}\mathbf{a}}_{\mathrm{Rel}} \to \min_{\substack{\mathbf{a} \in \mathbb{R}^n_+ \\ \|\mathbf{a}\|_1 = 1}}. \tag{4}$$

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 (4) 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 $\mathrm{Sim}(\mathbf{X})$ and $\mathrm{Rel}(\mathbf{X}, \boldsymbol{\phi})$ 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 matrix $\boldsymbol{\phi}$ for the function Rel

$$\mathbf{Q} = \left\{ \left| \mathrm{corr}(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j) \right| \right\}_{i,j=1}^{n}, \quad \mathbf{b} = \left\{ \left| \mathrm{corr}(\boldsymbol{\chi}_i, \boldsymbol{\phi}) \right| \right\}_{i=1}^{n}. \tag{5}$$

The problem (4) 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 (4) corresponds to the quality criteria $Q(\mathcal{A}|\mathbf{X}, \boldsymbol{\phi})$

$$\mathcal{A} = \operatorname*{arg\,max}_{\mathcal{A}' \subseteq \{1,\dots,n\}} Q(\mathcal{A}'|\mathbf{X}, \boldsymbol{\phi}) \Leftrightarrow \operatorname*{arg\,min}_{\mathbf{a} \in \mathbb{R}^n_+,\, \|\mathbf{a}\|_1 = 1} \left[\mathbf{a}^{\mathsf{T}}\mathbf{Q}\mathbf{a} - \alpha \cdot \mathbf{b}^{\mathsf{T}}\mathbf{a} \right]. \tag{6}$$

2.2 Multivariate QPFS

First of approach to apply the QPFS algorithm to the case, when $\mathbf{Y}$ is a matrix ($r > 1$) is to aggregate feature relevances throuch 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

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

ЭТО ПРИМЕР

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

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

We have three features and r targets, where first $r-1$ target are the same. 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.

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

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 $\boldsymbol{\chi}_1, \boldsymbol{\chi}_2$ predict the first component. The QPFS algorithm for $r = 2$ gives the solution $\mathbf{a} = [0.46, 0.31, 0.23]$. 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 $\mathbf{a} = [0.46, 0.25, 0.29]$. Here we lost the relevant feature $\boldsymbol{\chi}_2$ and select the redundant feature $\boldsymbol{\chi}_3$.

КОНЕЦ ПРИМЕРА

To take into account the dependencies in the columns of the matrix $\mathbf{Y}$ we extend the QPFS functional (4) to the multivariate case. We add the term $\text{Sim}(\mathbf{Y})$ and extend the term $\text{Rel}(\mathbf{X}, \mathbf{Y})$:

$$\alpha_1 \cdot \underbrace{\mathbf{a}_x^{\mathsf{T}} \mathbf{Q}_x \mathbf{a}_x}_{\text{Sim}(\mathbf{X})} - \alpha_2 \cdot \underbrace{\mathbf{a}_x^{\mathsf{T}} \mathbf{B} \mathbf{a}_y}_{\text{Rel}(\mathbf{X}, \mathbf{Y})} + \alpha_3 \cdot \underbrace{\mathbf{a}_y^{\mathsf{T}} \mathbf{Q}_y \mathbf{a}_y}_{\text{Sim}(\mathbf{Y})} \to \min_{\substack{\mathbf{a}_x \in \mathbb{R}^n_+ \, \|\mathbf{a}_x\|_1 = 1 \\ \mathbf{a}_y \in \mathbb{R}^r_+ \, \|\mathbf{a}_y\|_1 = 1}}. \tag{7}$$

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

$$\mathbf{Q}_x = \left\{ \left| \text{corr}(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j) \right| \right\}_{i,j=1}^n, \quad \mathbf{Q}_y = \left\{ \left| \text{corr}(\boldsymbol{\phi}_i, \boldsymbol{\phi}_j) \right| \right\}_{i,j=1}^r, \quad \mathbf{B} = \left\{ \left| \text{corr}(\boldsymbol{\chi}_i, \boldsymbol{\phi}_j) \right| \right\}_{\substack{i=1,\dots,n \\ j=1,\dots,r}}. \tag{8}$$

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

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

For the case $r = 1$ the proposed approach coincides with the original QPFS algorithm.

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.

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

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

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 $\mathbf{Y}$ matrix 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

$$\big\{ j : \mathrm{corr}(\boldsymbol{\chi}_j, \boldsymbol{\phi}_k) = 0, \, \forall k \in \{1, \dots, r\} \big\};$$

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

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

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

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

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

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

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

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