Multicorrelation

R. V. Isachenko, V. V. Strijov

Abstract: TBA

Keywords: TBA

1 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 (VIF) 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

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

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

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

3. non-**X**-correlated features, which are relevant to **Y**-correlated targets

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

4. **X**-correlated features, which are relevant to non-**Y**-correlated targets

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

5. **X**-correlated features, which are relevant to **Y**-correlated targets

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

Definition 1. The vectors $\boldsymbol{\chi}_1, \boldsymbol{\chi}_2 \in \mathbb{R}^m$ are called *δ-correlated* if

$$|\text{corr}(\boldsymbol{\chi}_1, \boldsymbol{\chi}_2)| \geq \delta.$$

Definition 2. The vectors $\boldsymbol{\chi}_1, \dots, \boldsymbol{\chi}_k \in \mathbb{R}^m$ are called *δ-multicorrelated* if

$$|\text{corr}(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j)| \geq \delta, \text{ for } i, j \in \{1, \dots, k\}.$$

Proposition 1. *The problem of extracting all δ-multicorrelated subsets from the given matrix are NP-complete.*

Proof. We show that the clique problem is reduced to the our problem which are NP-complete. Let consider the adjacency matrix of some graph. The vertices respond to columns of some matrix. The edges of this graph are pairs of the columns that are δ-correlated. The columns are δ-multicorrelated if all pairs of these columns are δ-correlated. In terms of the adjacency matrix it corresponds to a clique. □

Proposition 2. *The set of vectors which are δ-correlated with a vector $\boldsymbol{\nu} \in \mathbb{R}^m$ forms a cone:*

$$\mathsf{Cone}_\delta(\boldsymbol{\nu}) = \{\boldsymbol{\chi} \in \mathbb{R}^m : |corr(\boldsymbol{\chi}, \boldsymbol{\nu})| \geq \delta\}.$$

Proof. The proposition follows from the fact that

$$|\text{corr}(\boldsymbol{\chi}, \boldsymbol{\nu})| = |\text{corr}(\alpha\boldsymbol{\chi}, \boldsymbol{\nu})|, \text{ for } \alpha \geq 0.$$

Hence, the condition $\boldsymbol{\chi} \in \mathsf{Cone}_\delta(\boldsymbol{\nu})$ implies $\alpha\boldsymbol{\chi} \in \mathsf{Cone}_\delta(\boldsymbol{\nu})$. □

If vectors $\boldsymbol{\chi}_1, \dots, \boldsymbol{\chi}_k$ are δ-multicorrelated, then there is a vector $\boldsymbol{\nu}$ such that

$$\boldsymbol{\chi}_i \in \mathsf{Cone}_\delta(\boldsymbol{\nu}), \text{for } i = 1, \dots, k.$$

Since all vectors are pairwise δ-correlated, we could take any of $\boldsymbol{\chi}_i$ as the vector $\boldsymbol{\nu}$.

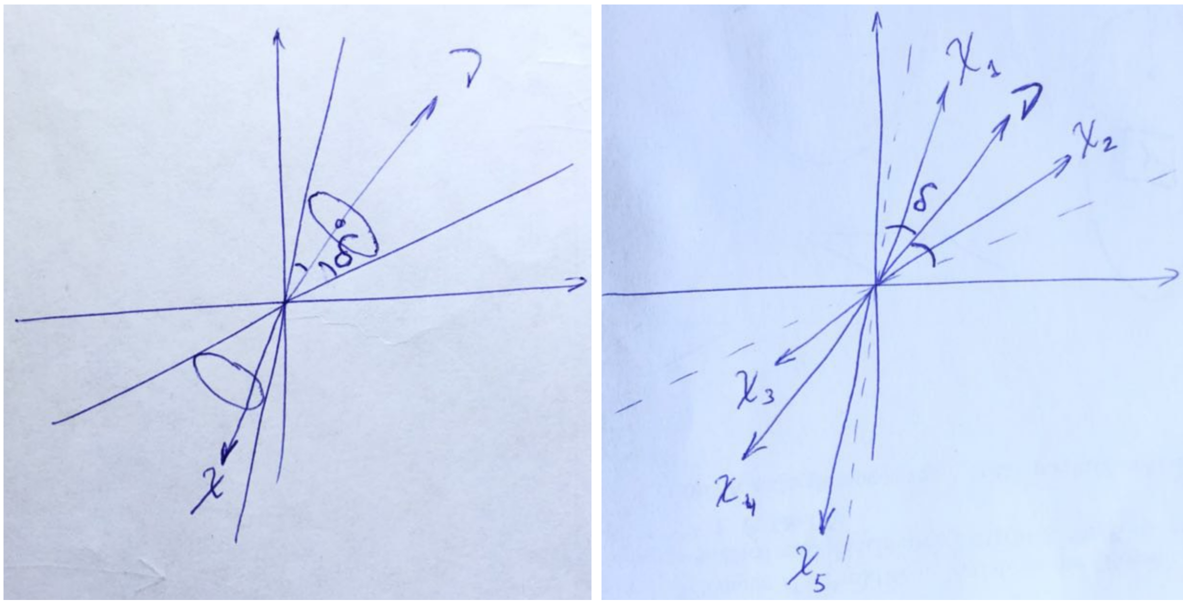

Figure 1: $\mathsf{Cone}_\delta(\boldsymbol{\nu})$ + δ-multicorrelation

There is a link between QPFS matrices $\mathbf{Q}_x$, $\mathbf{Q}_y$ and defined δ-multicorrelation. If we binarize the matrices and put 1's for the entries which are larger or equal to δ and 0's – otherwise, we will get the adjacency matrices of some graphs $G_{\mathbf{X}}$ and $G_{\mathbf{Y}}$. The edges in these graphs are pairs of vertices which are δ-correlated. The cliques in these adjacency matrices are the feature and the target subsets which are δ-multicorrelated. All vertices in $G_{\mathbf{X}}$ which are connected with the vertice i refer to features that lies in $\mathsf{Cone}_\delta(\boldsymbol{\chi}_i)$. Similarly, all vertices in $G_{\mathbf{Y}}$ which are connected with the vertice k refer to targets that lies in $\mathsf{Cone}_\delta(\boldsymbol{\nu}_k)$.

The binarized QPFS matrix $\mathbf{B}$ defines a bipartite graph $G_{\mathbf{XY}}$, where first part corresponds to the features and the second – to targets. In this notation the features that are non-relevant to targets are given by the vertices from the first part which are not connected to any vertex from the second part. We call features from the set $\mathsf{Cone}(\boldsymbol{\nu}_j)$ are relevant to the target $\boldsymbol{\nu}_j$.

We define two hypergraphs $H_{\mathbf{X}}$ and $H_{\mathbf{Y}}$. The hypergraphs are given by the set of vertices and the set of edges. There are n vertices in $H_{\mathbf{X}}$ and r vertices in $H_{\mathbf{Y}}$. The vertices respond to features and targets respectively. Each edge is given by a set of vertices that are δ-multicorrelated.

We propose the way to categorize all given features into five disjoint categories.

1. Non-relevant features. These features do not belong to any of the sets

$$\boldsymbol{\chi}_i \notin \mathsf{Cone}_\delta(\boldsymbol{\nu}_j) \text{ for } j = 1, \ldots, n.$$

   In the terms of QPFS algorithm for these features the corresponding rows of the matrix $\mathbf{B}$ contain only elements less than δ.

2. Non-correlated features, which are relevant to non-correlated targets.

$$
\begin{aligned}
&\exists j \in \{1, \ldots, n\} : \boldsymbol{\chi}_i \in \mathsf{Cone}_\delta(\boldsymbol{\nu}_j),\\
&\boldsymbol{\chi}_{i'} \notin \mathsf{Cone}_\mu(\boldsymbol{\chi}_i) \text{ for } i' \in \{1, \ldots, n\} : i' \neq i,\\
&\boldsymbol{\nu}_{j'} \notin \mathsf{Cone}_\lambda(\boldsymbol{\nu}_j) \text{ for } j' \in \{1, \ldots, n\} : j' \neq j.
\end{aligned}
$$

These features are isolated in the graph $G_{\mathbf{X}}$. There is an edge in the graph $G_{\mathbf{XY}}$ from these features to targets that are isolated in $G_{\mathbf{Y}}$.

3. Non-correlated features, which are relevant to correlated targets:

$$
\begin{aligned}
&\exists j \in \{1, \ldots, n\} : \boldsymbol{\chi}_i \in \mathsf{Cone}_\delta(\boldsymbol{\nu}_j),\\
&\boldsymbol{\chi}_{i'} \notin \mathsf{Cone}_\mu(\boldsymbol{\chi}_i) \text{ for } i' \in \{1, \ldots, n\} : i' \neq i,\\
&\exists j' \in \{1, \ldots, n\} : j' \neq j : \boldsymbol{\nu}_{j'} \in \mathsf{Cone}_\lambda(\boldsymbol{\nu}_j).
\end{aligned}
$$

These features are isolated in the graph $G_{\mathbf{X}}$. There is an edge in the graph $G_{\mathbf{XY}}$ from these features to targets that are not isolated in $G_{\mathbf{Y}}$.

4. Correlated features, which are relevant to non-correlated targets:

$$
\begin{aligned}
&\exists j \in \{1, \ldots, n\} : \boldsymbol{\chi}_i \in \mathsf{Cone}_\delta(\boldsymbol{\nu}_j),\\
&\exists i' \in \{1, \ldots, n\} : i' \neq i : \boldsymbol{\chi}_{i'} \in \mathsf{Cone}_\mu(\boldsymbol{\chi}_i),\\
&\boldsymbol{\nu}_{j'} \notin \mathsf{Cone}_\lambda(\boldsymbol{\nu}_j) \text{ for } j' \in \{1, \ldots, n\} : j' \neq j.
\end{aligned}
$$

These features are not isolated in the graph $G_{\mathbf{X}}$. There is an edge in the graph $G_{\mathbf{XY}}$ from these features to targets that are isolated in $G_{\mathbf{Y}}$.

5. Correlated features, which are relevant to correlated targets:

$$
\begin{aligned}
&\exists j \in \{1, \ldots, n\} : \boldsymbol{\chi}_i \in \mathsf{Cone}_\delta(\boldsymbol{\nu}_j),\\
&\exists i' \in \{1, \ldots, n\} : i' \neq i : \boldsymbol{\chi}_{i'} \in \mathsf{Cone}_\mu(\boldsymbol{\chi}_i),\\
&\exists j' \in \{1, \ldots, n\} : j' \neq j : \boldsymbol{\nu}_{j'} \in \mathsf{Cone}_\lambda(\boldsymbol{\nu}_j).
\end{aligned}
$$

These features are not isolated in the graph $G_{\mathbf{X}}$. There is an edge in the graph $G_{\mathbf{XY}}$ from these features to targets that are not isolated in $G_{\mathbf{Y}}$.