Comprehensive study of feature selection methods to solve multicollinearity problem according to evaluation criteria

A. M. Katrutsa[a,b,*], V. V. Strijov[a]

[a] *Moscow Institute of Physics and Technology, Institutskiy lane 9, Dolgoprudny city, 141700, Russian Federation*
[b] *Skolkovo Institute of Science and Technology, Nobel St., 3, Skolkovo, 143025, Russian Federation*

Abstract

This paper presents a comprehensive analysis of multicollinearity problem in data fitting. Data fitting is stated as a single-objective optimization problem where an objective function indicates the error of approximation the target vector with a some function of given features. The linear dependence between features means that the multicollinerity problem exists and leads to unstability and redundancy of the built model. These problems are addressed by introducing a feature selection method based on a quadratic programming approach. This approach takes into account the positions of the features and the target vector and select features according to relevance and similarity measures, which are defined by a user. Therefore, the built model is less redundant and more stable. To evaluate the quality of the proposed feature selection method and compare it with others we use different criteria to measure unstability and redundancy. In the experiments we compare proposed approach with other feature selection methods: LARS, Lasso, Ridge, Stepwise and Genetic algorithm. We show that the quadratic programming approach gives the best results according to considered criteria on the test and real data sets.

Keywords: data fitting, feature selection, multicollinearity, quadratic programming, evaluation criteria, test data sets

1. Introduction

This paper addresses the multicollinearity problem and proposes its comprehensive analysis. *Multicollinearity* is a strong correlation between features, which affect the target vector simultaneously. Due to multicollinearity the common methods of regression analysis like least squares build unstable models of excessive complexity. The formal definitions of model stability, complexity and redundancy are given in Section 5.

To treat multicollinearity problem feature selection methods are used. Most of previously proposed feature selection methods that solve multicollinearity problem are based on different heuristics [1, 2], greedy searches [3, 4] or regularization techniques [5, 6]. These approaches do not take into account the data set configuration and do not guarantee optimality of the obtained feature subset [7]. In constrast, we propose to use *quadratic programming approach* [8] to solve multicollinearity problem that corrects disadvantages mentioned above. This approach is based on two ideas: the first one is to represent features as some binary vector, and the second one is to define the feature subset quality criterion as quadratic form. The first term of the quadratic

[*] Corresponding author
Email address: aleksandr.katrutsa@phystech.edu (A. M. Katrutsa)

form is pairwise feature similarities and the linear term is feature relevances. Therefore, we can state feature selection problem with the quadratic objective and boolean vector domain. The measures of feature similarities and relevances are problem-dependent and have to be defined by user before performing feature selection. These measures have to take into account the data set configuration to remove redundant, noisy and multicollinear features, but select features which are significant for target vector approximation. We consider the correlation coefficient [9] and the mutual information [10] between features as measures of feature similarities and between features and target vector as measure of feature relevances. These measures give positive semidefinite quadratic form, and to get *convex optimization problem* we need to relax binary domain to continuous one. After this relaxation we have *convex optimization problem*, which can be efficiently solved by state-of-the-art solvers, for example from CVX, a package for specifying and solving convex programs package [11, 12]. To return from the continuous solution to the binary one, we need a *significance threshold*, which defines what features are selected. If one would like to use a feature similarity function that does not give positive semidefinite matrix, then optimization problem is not convex, and convex relaxation is required. In this case, the authors propose the use semidefinite programming relaxation [13]. Such feature similarity functions are out of scope of this paper. In addition, the proposed approach gives simple visualization of the feature weights in the target vector approximation. This visualization helps tuning the threshold in the most appropriate way for user.

We carry out experiments on special test data sets generated according to the procedure proposed in [7]. These data sets demostrate different cases of multicollinearity between features and correlation between features and target vector. Experimets show that the proposed approach outperforms other considered feature selection methods on every type of test data sets. Also quadratic programming feature selection shows better quality on test and real data sets according to various evaluation criteria simultaneously in constrast to other feature selection methods.

The main contributions of this paper are:

- We address multicollinearity problem with quadratic programming approach and investigate its property.
- We demonstrate performance of the quadratic programming feature selection method on the test data sets according to various criteria.
- We compare the proposed feature selection method with others on test and real data sets and show that it gives the better feature subset than other methods. The feature subset quality are measured by external criteria.

Related works. Previously, authors propose different strategies to detect multicollinearity problem and approaches to solve this problem [14, 15, 16]. One way to solve multcollinearity problem is to use feature selection methods [16]. They are based on some score function which estimates quality of feature subset or some heuristic sequaential search procedure. This paper considers feature selection methods, which are based on scoring function, like LARS [17], Lasso [18], Ridge [6], Elastic Net [5], and which are based on the sequential search like Stepwise [19] and Genetic algorithm [20].

2. Feature Selection Problem Statement

Let $\mathbf{X} = [\boldsymbol{\chi}_1, \dots, \boldsymbol{\chi}_n] \in \mathbb{R}^{m \times n}$ be the design matrix, where $\boldsymbol{\chi}_j \in \mathbb{R}^m$ is an j-th feature. Let $\mathbf{y} \in \mathbb{R}^m$ be the target vector. Denote by $\mathcal{J} = \{1, \dots, n\}$ a feature index set. Let $\mathcal{A} \subseteq \mathcal{J}$ be a feature index subset. The data fitting problem is to find a parameter vector $\mathbf{w}^* \in \mathbb{R}^n$ such that:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathbb{R}^n} S(\mathbf{w}, \mathcal{A}|\mathbf{X}, \mathbf{y}, \mathbf{f}), \tag{1}$$

where S is an error function, which validates the quality of any parameter vector $\mathbf{w}$ and corresponding feature index subset $\mathcal{A}$ with given design matrix $\mathbf{X}$, target vector $\mathbf{y}$ and a function $\mathbf{f}$. The function $\mathbf{f}$ approximates the target vector $\mathbf{y}$.

This study uses linear function:

$$\mathbf{f}(\mathbf{X}, \mathcal{A}, \mathbf{w}) = \mathbf{X}_{\mathcal{A}}\mathbf{w},$$

where $\mathbf{X}_{\mathcal{A}}$ is the reduced design matrix which consists of only the feature with indices from the set $\mathcal{A}$, and quadratic error function

$$S(\mathbf{w}, \mathcal{A}|\mathbf{X}, \mathbf{y}, \mathbf{f}) = \|\mathbf{f}(\mathbf{X}, \mathcal{A}, \mathbf{w}) - \mathbf{y}\|_2^2. \tag{2}$$

The features $\boldsymbol{\chi}_j, j \in \mathcal{J}$ are supposed to be noisy, irrelevant or multicollinear that leads to additional error in estimation of the optimum vector $\mathbf{w}^*$ and unstability of this vector. One can use feature selection methods to remove named features from design matrix $\mathbf{X}$. The feature selection procedure reduces dimensionality of problem (1) and improves stability of the optimum vector $\mathbf{w}^*$. The feature selection problem is

$$\mathcal{A}^* = \arg\min_{\mathcal{A}\subseteq\mathcal{J}} Q(\mathcal{A}|\mathbf{X}, \mathbf{y}), \tag{3}$$

where $Q : \mathcal{A} \to \mathbb{R}$ is a quality criterion, which validates the quality of some selected feature index subset $\mathcal{A} \subseteq \mathcal{J}$. Problem (3) does not necessarily require any estimation of the optimum parameter vector $\mathbf{w}^*$. It uses relations between the features $\boldsymbol{\chi}_j, j = 1, \dots, n$ and the target vector $\mathbf{y}$.

Let $\mathbf{a} \in \mathbb{B}^n = \{0, 1\}^n$ be an indicator vector such that $a_j = 1$ if and only if $j \in \mathcal{A}$. So the problem (3) can be rewritten in the following form:

$$\mathbf{a}^* = \arg\min_{\mathbf{a}\in\mathbb{B}^n} Q(\mathbf{a}|\mathbf{X}, \mathbf{y}), \tag{4}$$

where $Q : \mathbb{B}^n \to \mathbb{R}$ is another form of the criterion Q with domain $\mathbb{B}^n$. The vector $\mathbf{a}^*$ and the index set $\mathcal{A}^*$ are equivalent in the following sense:

$$a_j^* = 1 \Leftrightarrow j \in \mathcal{A}^*, \ j \in \mathcal{J}. \tag{5}$$

2.1. Multicollinearity problem

In this subsection we give formal definition of multicollinearity phenomenon and special cases. Assume that features $\boldsymbol{\chi}_j$ and target vector $\mathbf{y}$ are normalized:

$$\|\mathbf{y}\|_2 = 1 \text{ and } \|\boldsymbol{\chi}_j\|_2 = 1, \ j \in \mathcal{J}. \tag{6}$$

Consider active index subset $\mathcal{A} \subseteq \mathcal{J}$.

Definition 2.1 The features with indices from the set $\mathcal{A}$ are called *multicollinear* if there exist the index j, the coefficients a_k, the index $k \in \mathcal{A} \setminus j$ and sufficiently small positive number $\delta > 0$ such that

$$\left\| \boldsymbol{\chi}_j - \sum_{k\in\mathcal{A}\setminus j} \lambda_k \boldsymbol{\chi}_k \right\|_2^2 < \delta. \tag{7}$$

The smaller δ is, the higher *degree of multicollinearity*.

The particular case of this definition is the following one.

Definition 2.2 Let the features indexed i, j be *correlated* if there exists sufficiently small positive number $\delta_{ij} > 0$ such that:

$$\|\boldsymbol{\chi}_i - \boldsymbol{\chi}_j\|_2^2 < \delta_{ij}. \tag{8}$$

From this definition it follows that $\delta_{ij} = \delta_{ji}$. Inequalities (7) and (8) are identical if $a_k = 0 \; k \neq j$ and $a_k = 1 \; k = j$.

Definition 2.3 Feature $\boldsymbol{\chi}_j$ is called *correlated with the target vector* $\mathbf{y}$ if there exists sufficiently small positive number $\delta_j > 0$ such that

$$\|\mathbf{y} - \boldsymbol{\chi}_j\|_2^2 < \delta_j.$$

3. Quadratic Optimization Approach to Multicollinearity Problem

The paper [7] shows that none of he considered feature selection methods (LARS, Lasso, Ridge, Stepwise and Genetic algorithm) solve the problem (1) and give stable, accurate and nonredundant model simultaneously. Therefore, we propose the quadratic programming approach to solve multicollinearity problem.

The main idea of the proposed approach is to minimize the number of similar features and maximize the number of relevant features. To formalize this idea we represent the criterion Q from the problem (4) in the form of quadratic function:

$$Q(\mathbf{a}) = \mathbf{a}^{\mathsf{T}}\mathbf{Q}\mathbf{a} - \mathbf{b}^{\mathsf{T}}\mathbf{a}, \tag{9}$$

where $\mathbf{Q} \in \mathbb{R}^{n \times n}$ is a matrix of pairwise features similarities, $\mathbf{b} \in \mathbb{R}^n$ is a vector of features relevances to the target vector.

To indicate the matrix $\mathbf{Q}$ and vector $\mathbf{b}$ computation approach, introduce functions Sim and Rel:

$$\begin{aligned} &\text{Sim: } \mathcal{J} \times \mathcal{J} \to [0, 1], \\ &\text{Rel: } \mathcal{J} \to [0, 1]. \end{aligned} \tag{10}$$

These functions are problem-dependent, defined by user before performing feature selection and indicate the way to measure feature similarities (Sim) and relevance to the target vector (Rel). To highlight the dependence quadratic programming feature selection method on similarity and relevance functions, introduce the following definition.

Definition 3.1 Let QP(Sim, Rel) be a feature selection method, which solves the follwing optimization problem:

$$\mathbf{a}^* = \arg\min_{\mathbf{a} \in \mathbb{B}^n} \mathbf{a}^{\mathsf{T}}\mathbf{Q}\mathbf{a} - \mathbf{b}^{\mathsf{T}}\mathbf{a}, \tag{11}$$

where the matrix $\mathbf{Q}$ is computed by the function Sim:

$$\mathbf{Q} = [q_{ij}] = \text{Sim}(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j). \tag{12}$$

and the vector $\mathbf{b}$ is computed by the function Rel:

$$\mathbf{b} = [b_i] = \text{Rel}(\boldsymbol{\chi}_i). \tag{13}$$

Below we provide examples of functions Sim and Rel, which illustrate the proposed approach.

3.1. Correlation coefficient

The similarities between features $\boldsymbol{\chi}_i$ and $\boldsymbol{\chi}_j$ can be computed with the Pearson correlation coefficient [9]. The Pearson correlation coefficient is defined as:

$$\rho_{ij} = \frac{\mathrm{Cov}(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j)}{\sqrt{\mathrm{Var}(\boldsymbol{\chi}_i)\mathrm{Var}(\boldsymbol{\chi}_j)}},$$

where $\mathrm{Cov}(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j)$ is a covariance between features $\boldsymbol{\chi}_i$ and $\boldsymbol{\chi}_j$, $\mathrm{Var}(\cdot)$ is a variance of a feature. The sample correlation coefficient is defined as

$$\hat{\rho}_{ij} = \frac{\sum\limits_{k=1}^{m}(\boldsymbol{\chi}_{ik} - \overline{\boldsymbol{\chi}}_i)(\boldsymbol{\chi}_{jk} - \overline{\boldsymbol{\chi}}_j)}{\sqrt{\sum\limits_{k=1}^{m}(\boldsymbol{\chi}_{ki} - \overline{\boldsymbol{\chi}}_i)^2 \sum\limits_{k=1}^{m}(\boldsymbol{\chi}_{kj} - \overline{\boldsymbol{\chi}}_j)^2}}. \tag{14}$$

In this case the elements of the matrix $\mathbf{Q} = [q_{ij}]$ are equal to the absolute values of the corresponding sample correlation coefficients:

$$q_{ij} = \mathrm{Sim}(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j) = |\hat{\rho}_{ij}| \tag{15}$$

and the elements of the vector $\mathbf{b} = [b_i]$ are equal to absolute values of the sample correlation coefficient between feature $\boldsymbol{\chi}_i$ and the target vector $\mathbf{y}$:

$$b_i = \mathrm{Rel}(\boldsymbol{\chi}_i) = |\hat{\rho}_{iy}|. \tag{16}$$

It means that we want to minimize the number of correlated features and maximize the number of features correlated to the target vector.

3.2. Mutual information

One more feature similarity measure is based on the mutual information concept [10]. The mutual information between features $\boldsymbol{\chi}_i$ and $\boldsymbol{\chi}_j$ is defined as

$$I(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j) = \int\int p(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j) \log \frac{p(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j)}{p(\boldsymbol{\chi}_i)p(\boldsymbol{\chi}_j)} d\boldsymbol{\chi}_i d\boldsymbol{\chi}_j. \tag{17}$$

The sample mutual information is calculated based on estimation of the probability distribution in the equation (17). In this case the elements of the matrix $\mathbf{Q} = [q_{ij}]$ are equal to the value of the corresponding sample mutual information:

$$q_{ij} = \mathrm{Sim}(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j) = I(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j)$$

and the elements of the vector $\mathbf{b} = [b_i]$ are equal the sample mutual information of every feature and the target vector:

$$b_i = \mathrm{Rel}(\boldsymbol{\chi}_i) = I(\boldsymbol{\chi}_i, \mathbf{y}).$$

3.3. Normalized feature significance

The correlation coefficient (14) and mutual information (17) do not directly present the feature relevance. To take into account features relevance we propose to use the normalized significance of the features estimated by t-test. To select relevant features, state the following hypothesis testing problem for every $j-th$ feature:

$$\begin{aligned} H_0 &: w_j = 0, \\ H_1 &: w_j \neq 0. \end{aligned} \tag{18}$$

The obtained p-value p_j shows the j-th feature relevance in the target vector approximation. If $p_j < 0.05$, than we reject the H_0 hypothesis and suppose that the corresponding j-th element of the parameter vector w_j is not zero.

Definition 3.2 Let $\hat{p}_j$ be a *normalized feature significance* for the j-th feature, $j \in \mathcal{J}$:

$$\hat{p}_j = 1 - \frac{p_j}{\sum\limits_{k=1}^{n} p_k}.$$

Thus, to represent the feature relevance we propose to use in (13) normalized feature significance:

$$b_j = \text{Rel}(\boldsymbol{\chi}_j) = \hat{p}_j. \tag{19}$$

3.4. Convex representation of the problem (11)

Quadratic programming approach to multicollinearity problem leads to the problem (4), which is NP-hard due to boolean domain. Therefore, we need to approximate it with the convex oprimization problem to solve it efficiently.

Assume that function Sim gives the positive semidefinite matrix $\mathbf{Q}$, then the quadratic form (9) is the convex function. To represent problem (11) in the convex form we have to replace nonconvex set $\mathbb{B}^n$ with the convex one. The natural way for this representation is to use the convex hull of the set $\mathbb{B}^n$:

$$\text{Conv}(\mathbb{B}^n) = [0, 1]^n.$$

Now the problem (11) is approximated by the following *convex optimization problem*:

$$\begin{aligned} \mathbf{z}^* &= \underset{\mathbf{z} \in [0,1]^n}{\arg\min}\, \mathbf{z}^{\mathsf{T}} \mathbf{Q} \mathbf{z} - \mathbf{b}^{\mathsf{T}} \mathbf{z} \\ \text{s.t. } \|\mathbf{z}\|_1 &\leq 1. \end{aligned} \tag{20}$$

We add the constraint to show that $\mathbf{z}^*$ can be treated as a vector non-normalized probabilities for every feature in active set $\mathcal{A}^*$.

To return from the continuous vector $\mathbf{z}^*$ to the boolean vector $\mathbf{a}^*$ and consequently to the active set $\mathcal{A}^*$ (see (5)), we use a *significance threshold* τ.

Definition 3.3 Let τ be a significance threshold such that $z_j^* > \tau$ if and only if $a_j^* = 1$ and $j \in \mathcal{A}^*$.

Tuning τ is problem-dependent and based on the appropriate error rate, number of selected features and values of evaluation criteria. One has to check some range of τ to get the most appropriate one for considered problem. In Section 6 we show examples of tuning τ.

4. Test Data Sets

To test the proposed quadratic programming approach in the case of extremely feature correlation we use synthetic test data sets from the paper [7]. These data sets demonstrate the performance of feature selection methods in the treating multicollinearity problem. Below we provide summary of the proposed test data sets.

Definition 4.1 Let *inadequate and correlated data set* be a data set that consists of the correlted features, which are otrhogonal to the target vector. Fig. 1a) demonstrates configuration of such data set.

Definition 4.2 Let *adequate and random data set* be a data set that consists of the random features with the single feature which approximates the target vector. Fig. 1b) demonstrates configuration of such data set.

Definition 4.3 Let *adequate and redundant data set* be a data set that consists of the features with are correlated to the target vector. Fig. 1c) demonstrates configuration of such data set.

Definition 4.4 Let *adequate and correlated data set* be a data set that consists of the orthogonal features and features, corelated to the orthogonal ones; the taget vector is a sum of two orthogonal features. Fig. 1d) demonstrates configuration of such data set.

Performance of the considered feature selection methods are compared according to various evaluation criteria which are provided in the next section.

5. Evaluation Criteria

To evaluate quality of the selected feature subset and compare considered feature selection methods we use the following criteria used in papers [21, 22].

Variance inflation factor. To diagnose multicollinearity, the paper [21] uses the variance inflation factor VIF_j. The VIF_j shows a linear dependence between the j-th feature and the other features. To compute VIF_j estimate the parameter vector $\mathbf{w}^*$ according to the problem (1) assuming $\mathbf{y} = \boldsymbol{\chi}_j$ and extracting j-th feature from the index set $\mathcal{A} = \mathcal{A} \setminus j$. The VIF_j is computed with the following equation:

$$\text{VIF}_j = \frac{1}{1 - R_j^2},$$

where $R_j^2 = 1 - \frac{\text{RSS}_j}{\text{TSS}_j}$ is the coefficient of determination and

$$\text{RSS}_j = \|\boldsymbol{\chi}_j - \mathbf{X}_{\mathcal{A}}\mathbf{w}^*\|_2^2, \qquad \text{TSS}_j = \sum_{i=1}^{m}(\chi_{ji} - \overline{\boldsymbol{\chi}}_j)^2, \quad \overline{\boldsymbol{\chi}}_j = \frac{1}{m}\sum_{i=1}^{m}\chi_{ji}.$$

The paper [21] states that if $\text{VIF}_j \gtrsim 5$ then the associated element of the vector $\mathbf{w}^*$ is poorly estimated because of multicollinearity. Denote by VIF the maximum value of VIF_j for all $j \in \mathcal{A}$:

$$\text{VIF} = \max_{j \in \mathcal{A}} \text{VIF}_j.$$

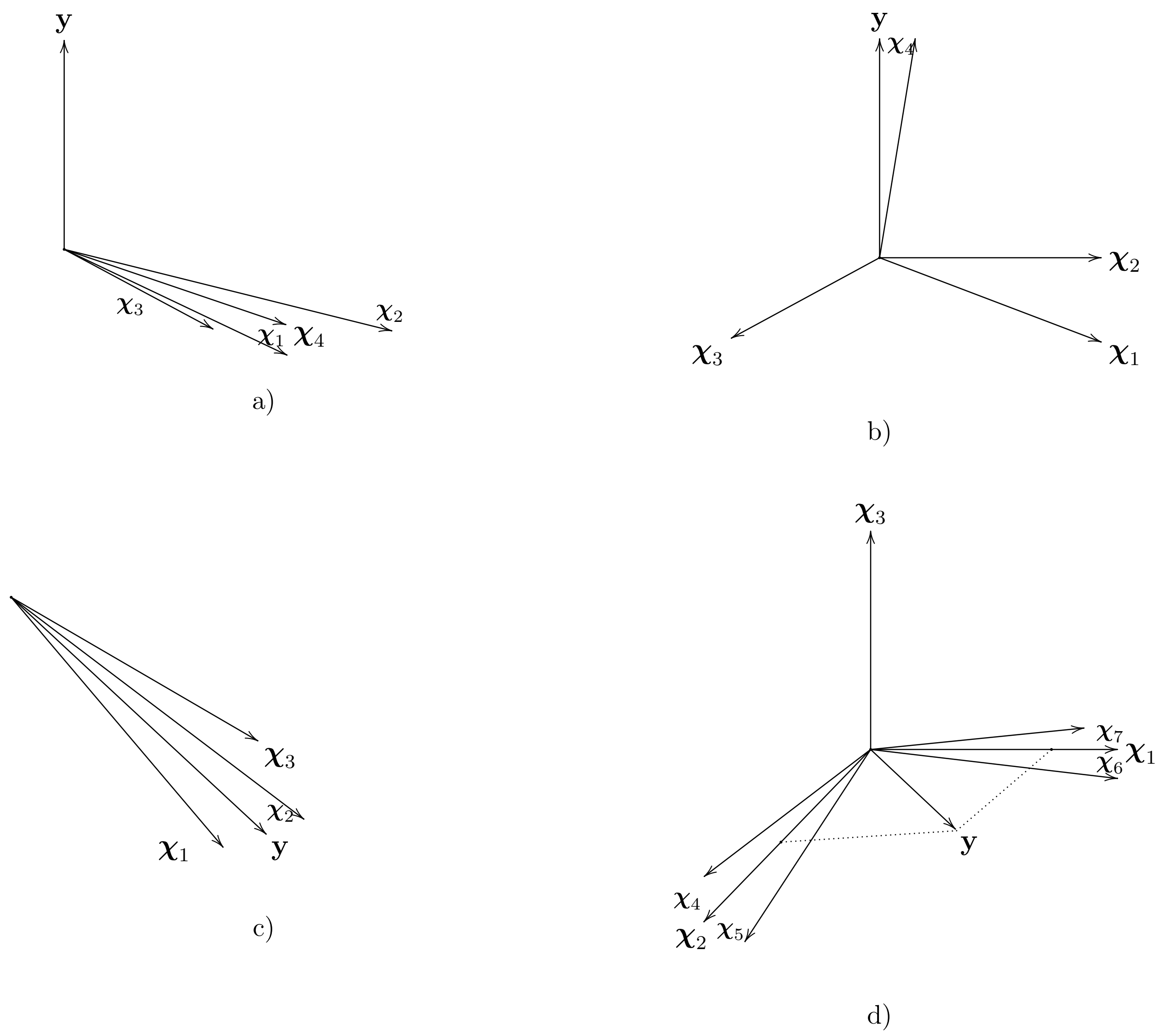


Figure 1: Synthetic test data sets configuration: a) inadequate and correlated, b) adequate and random, c) adequate and redundant, d) adequate and correlated.

Stability. To estimate the stability R of the parameter $\mathbf{w}$ estimation based on the selected feature subset $\mathcal{A}$, we use the logarithm of the inverse condition number of matrix $\mathbf{X}^\mathsf{T}\mathbf{X}$:

$$R = \ln \frac{\lambda_{\min}}{\lambda_{\max}},$$

where the $\lambda_{\max}$ and $\lambda_{\min}$ are the maximum and minimum non-zero eigenvalues of the matrix $\mathbf{X}^\mathsf{T}\mathbf{X}$. The larger R is, the more stable parameter estimation.

Complexity. To measure complexity C of the selected feature subset $\mathcal{A}^*$ we use the cardinality of this subset, i.e.

$$C = |\mathcal{A}^*|.$$

The less complexity is, the better selected subset.

Mallow's C_p. The Mallow's C_p criterion [23] trades off the residual norm S (2) and the number of features p. The Mallow's C_p defined as

$$C_p = \frac{r_p}{r} - m + 2p,$$

where S_p is similar to S, but computed with $p = |\mathcal{A}|$ features only. In terms of this criterion the smaller C_p is, the better feature subset.

BIC. Information criterion BIC [24] defined as

$$\text{BIC} = r + p \log m.$$

The smaller value of BIC is, the better model fits the target vector. Considered criteria are summarized in the table 1.

Table 1: A list of the criteria to evaluate the selected feature subset

Name	*Formula*	*Meaning*
VIF	$\text{VIF} = \max\limits_{j \in \mathcal{A}} \frac{1}{1-R_j^2}$	Indicator of the the multicollinear features existence
Stability	$R = \ln \frac{\lambda_{\min}}{\lambda_{\max}}$	An indicator of the model stability
Complexity	$C = \lvert\mathcal{A}^*\rvert$	The number of the selected features
Mallow's C_p	$C_p = \frac{r_p}{r} - m + 2p$	A trade-off between accuracy and number of features
BIC	$\text{BIC} = r + p \log m$	A trade-off between residues norm and number of features

6. Computational experiment

In this section we provide the experiments on the synthetic and real data sets to show the performance of the proposed approach in multicollinearity problem. The source code can be found at Github[1].

6.1. Data

We use the synthetic test data sets generated according to the procedure proposed in [7] to investigate performance of the considered methods from the multicollinearity problem point of view. The test data sets configurations are described in Section 4. The parameters of the test data sets are the following: number of objects $m = 1000$, number of features $n = 50$. Also we use the publicly available real dataset of diesel fuels NIR spectra [25].

[1] https://github.com/amkatrutsa/QPFeatureSelection

6.2. Comparison with other feature selection methods

Tables 2, 3, 4 and 5 show that the proposed approach is appropriate for every test data set configuration described in Section 4 in contrast with other feature selection methods. The methods are sorted in the descending order: the higher method is, the better. The choice of the functions Sim and Rel for every data set configuration is based on the dependendces between features and target vector. The significance threshold τ is chosen for every data set separately. The dash in table cell indicates that the number of selected features is zero.

Table 2: Evaluation criteria for the inadequate correlated data set — Fig. 1a)

Method	C_p	RSS	R	VIF	BIC
QP(ρ, ρ) ($\tau = 8 \cdot 10^{-5}$)	-997	—	—	—	—
LARS	-997	—	—	—	—
Genetic	-997	—	—	—	—
Lasso	-997	1	-6.57	16.6	310.48
Ridge	-997	1	-6.69	16.6	346.39
Stepwise	-997	1.68	-6.69	16.6	347.01
Elastic Net	-997	1	-6.58	16.6	310.48

Table 3: Evaluation criteria for the adequate and random data sets — Fig. 1b)

Method	C_p	RSS	R	VIF	BIC
QP(ρ, ρ) ($\tau = 10^{-4}$)	-997	$1.2 \cdot 10^{-9}$	0	0.24	6.9
Lasso	$7 \cdot 10^{6}$	$8.50 \cdot 10^{-4}$	0	0.25	6.9
Elastic Net	$8.76 \cdot 10^{-4}$	$8.76 \cdot 10^{-4}$	0	0.25	6.9
Ridge	$7.97 \cdot 10^{9}$	0.97	0	0.25	7.88
LARS	-997	$1.3 \cdot 10^{-10}$	-0.78	0.32	8.29
Genetic	-997	$1.36 \cdot 10^{-10}$	-3.31	0.9	52.5
Stepwise	-997	$1.33 \cdot 10^{-10}$	-3.36	0.89	53.88

Table 4: Evaluation criteria for the adequate and redundant data set — Fig. 1c)

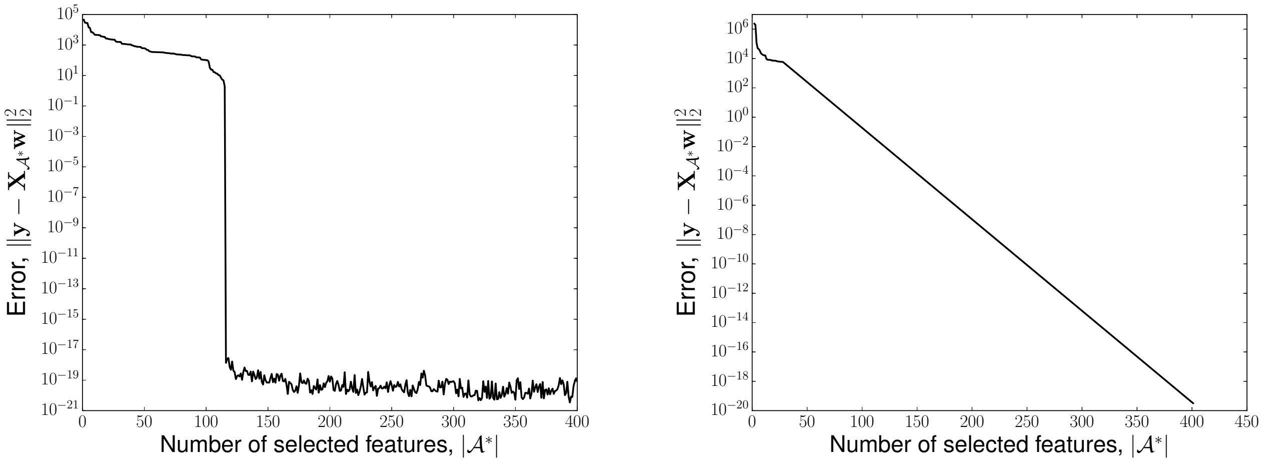
Method	C_p	RSS	R	VIF	BIC
QP(ρ, ρ) ($\tau = 10^{-2}$)	-997	$8.5 \cdot 10^{-11}$	0	0.25	6.9
Lasso	$5.16 \cdot 10^{8}$	$8.5 \cdot 10^{-4}$	0	0.24	6.9
Ridge	$5.9 \cdot 10^{11}$	0.97	-27.13	$2.9 \cdot 10^{9}$	346.36
Elastic Net	$5.16 \cdot 10^{8}$	$8.5 \cdot 10^{-4}$	-25.01	$2.5 \cdot 10^{9}$	41.45
Genetic	-997	$1.67 \cdot 10^{-12}$	-27.11	$2.87 \cdot 10^{9}$	345.39
Stepwise	-997	$1.73 \cdot 10^{-12}$	-27.13	$2.9 \cdot 10^{9}$	345.39
LARS	-997	$1.65 \cdot 10^{-12}$	-27.13	$2.9 \cdot 10^{9}$	345.39

Now we provide the similar analysis for NIR spectra of diesel fuel dataset in Fig. 2, where we compare dependence of residual norm on the number of the selected features based on correlation coefficient and mutual information similarity measures. We have to provide this comparison because we do not know the configuration of real data set in constrast with the test data sets. Fig. 2 shows that correlation coefficient similarity measure is better than mutual information to identify

Table 5: Evaluation criteria for the adequate and correlated data set — Fig. 1d)

Method	C_p	RSS	R	VIF	BIC
QP(ρ, $\hat{p}$) ($\tau = 10^{-7}$)	$9.1 \cdot 10^5$	$7.5 \cdot 10^{-25}$	0	0.63	13.8
Stepwise	$9.4 \cdot 10^5$	$8.8 \cdot 10^{-25}$	0	0.63	13.82
Genetic	$4.95 \cdot 10^7$	$2.93 \cdot 10^{-23}$	0	0.63	13.81
Ridge	$1.8 \cdot 10^{30}$	0.95	−36.8	$8.65 \cdot 10^{16}$	152.92
LARS	10^{30}	0.38	−67.87	10^{20}	108.15
Lasso	$1.73 \cdot 10^{27}$	$9.2 \cdot 10^{-4}$	−36.83	10^{17}	150.59
Elastic Net	$1.7 \cdot 10^{27}$	$9.2 \cdot 10^{-4}$	−36.83	10^{17}	150.59

the minimum number of features which give appropriate quality.



a) b)

Figure 2: Dependence of resudual norm on the number of selected features a) QP(ρ, ρ) for correlation coefficient and b) QP(I, I), for mutual information

Table 6 compares the considered approach with other feature selection methods on the NIR spectra of diesel fuel. This table shows that quadratic programming approach is comparable with other considered feature selection methods.

Table 6: Evaluation criteria for the diesel NIR spectra dataset

Method	C_p	RSS	R	VIF	BIC
QP (ρ, ρ) ($\tau = 10^{-9}$)	−110	$1.37 \cdot 10^{-18}$	−25.7	$6.43 \cdot 10^6$	548.38
Genetic	−110.88	$7.68 \cdot 10^{-30}$	−24	$8.13 \cdot 10^5$	534.19
LARS	$3.22 \cdot 10^{21}$	$2.07 \cdot 10^{-7}$	−28.3	$7.94 \cdot 10^7$	529.47
Lasso	$2.5 \cdot 10^{28}$	1.61	−27.72	$1.03 \cdot 10^{21}$	1712.92
ElasticNet	$2.51 \cdot 10^{28}$	1.61	−27.72	$1.03 \cdot 10^{21}$	1712.92
Stepwise	$3.66 \cdot 10^{29}$	23.56	−36.78	$1.94 \cdot 10^{22}$	1919.23
Ridge	$1.59 \cdot 10^{28}$	1.02	−36.22	$1.07 \cdot 10^{22}$	$1.79 \cdot 10^3$

6.3. Tuning significance threshold τ

This paragraph provides methodology of tuning significance threshold τ through simple visualization and choosing the most important criterion from Table 1. The general approach to tune significance threshold τ is to plot dependence of some evaluation criterion or criteria on some range of τ. These plots demonstrate what value of τ is more appropriate for considered data set. Fig. 3 and 4 show dependence number of features and error function on the significance threshold τ. We use these plots to select τ for tables 2, 3, 4 and 5 in subsection 6.2.

To tune τ we use the correlation coefficient (15) to generate matrix $\mathbf{Q}$, and correlation between features and target vector (16) or normalized feature significance (19) to generate linear term $\mathbf{b}$. Fig. 3 shows the number of selected features versus the chosen threshold τ for every kind of synthetic test data sets. Fig. 3a) shows that all features have the same and very small weights, which means that these features are irrelevant to target vector approximation. Fig. 3c) shows that all features have the same weights, but in contrast with Fig. 3a) these weights are much bigger, which means that all features are relevant and any feature can be selected to approximate target vector. Fig. 3d) shows that most features are irrelevant or redundant and only small number of features, namely 2, is relevant to precise target vector approximation.

Fig. 4 shows dependence of the error function S on significance threshold τ. Fig 4a) shows that for any τ the error is constant, which confirms the interpretation of Fig. 3a) about features irrelevance.

Fig. 4b) shows slow increasing of error function with increasing τ. In case of random data set this dependence means that more features are, lower aproximation error. Combined Fig. 3b) and 4b) we conclude that significance threshold $\tau \approx 10^{-4}$ is the best trade-off between number of features and approximation error.

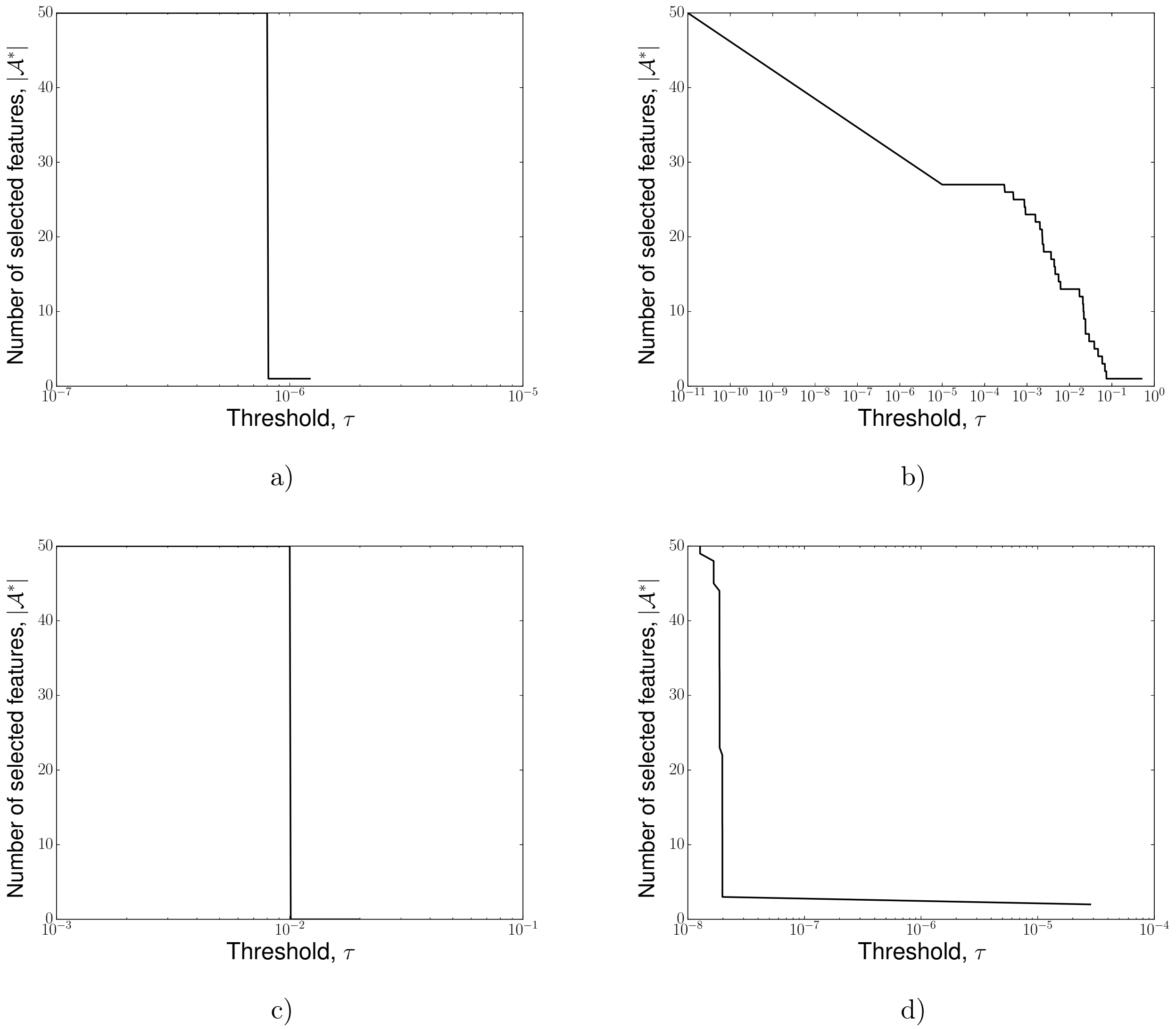
Fig. 4c) shows that error decreases significantly in case of exact $\tau = 10^{-2}$. Compare with Fig. 3c) and conclude that the single feature is enough to approximate target vector, which is consistent with the known adequate and redundant data set configuration, see Fig. 1c).

Fig. 4d) shows that most features lead to error oscillation (near $\tau = 10^{-8}$) and after selection 2 relevant features approximation error is stable low. This dependence is consistent with known adequate and correlated data set configuration, see Fig. 1d).

Thus, the described methodology shows that tuning significance threshold τ and quadratic programming feature selection give reasonable result in feature selection problem. The provided plots demonsrate ability of the proposed approach to extract considered patterns of multicollinearity from test data sets. The main reason of this ability is the choice of the functions Sim and Rel, which give the appropriate estimations of the features similarity and relevance.

7. Conclusion

This study addresses multicollinearity problem from the quadratic programming point of view. Quadratic programming approach gives reasonable methodology to investigate features relevance and redundancy. The proposed approach is tested on the syntehtic test data sets with special configurations of features and target vector. These configurations demonstrate different cases of the multicollinearity problem. Under multicollinearity conditions the quadratic programming feature selection method outperforms other feature selection methods like LARS, Lasso, Stepwise, Ridge and Genetic algorithm on the considered test and real data sets. Also, we compare performance of the proposed approach with other feature selection methods according to various evaluation criteria and show that the proposed approach gives higher quality feature subsets than other methods.



a)

b)

c)

d)

Figure 3: Dependence the cardinality of the active index set $\mathcal{A}$ on the threshold τ for: a) inadequate correlated data set, b) adequate random data set, c) adequate redundant data set, d) adequate correlated data set.

8. Acknowledgment

This study was supported by RFBR project 16-37-00485.

References

[1] R. Leardi, Genetic algorithms in chemometrics and chemistry: a review, Journal of chemometrics 15 (7) (2001) 559–569.

[2] B. Oluleye, L. Armstrong, J. Leng, D. Diepeveen, A genetic algorithm-based feature selection, British Journal of Mathematics & Computer Science (2014) In–Press.

[3] L. Ladha, T. Deepa, Feature selection methods and algorithms, International Journal on Computer Science and Engineering 3 (5) (2011) 1787–1797.

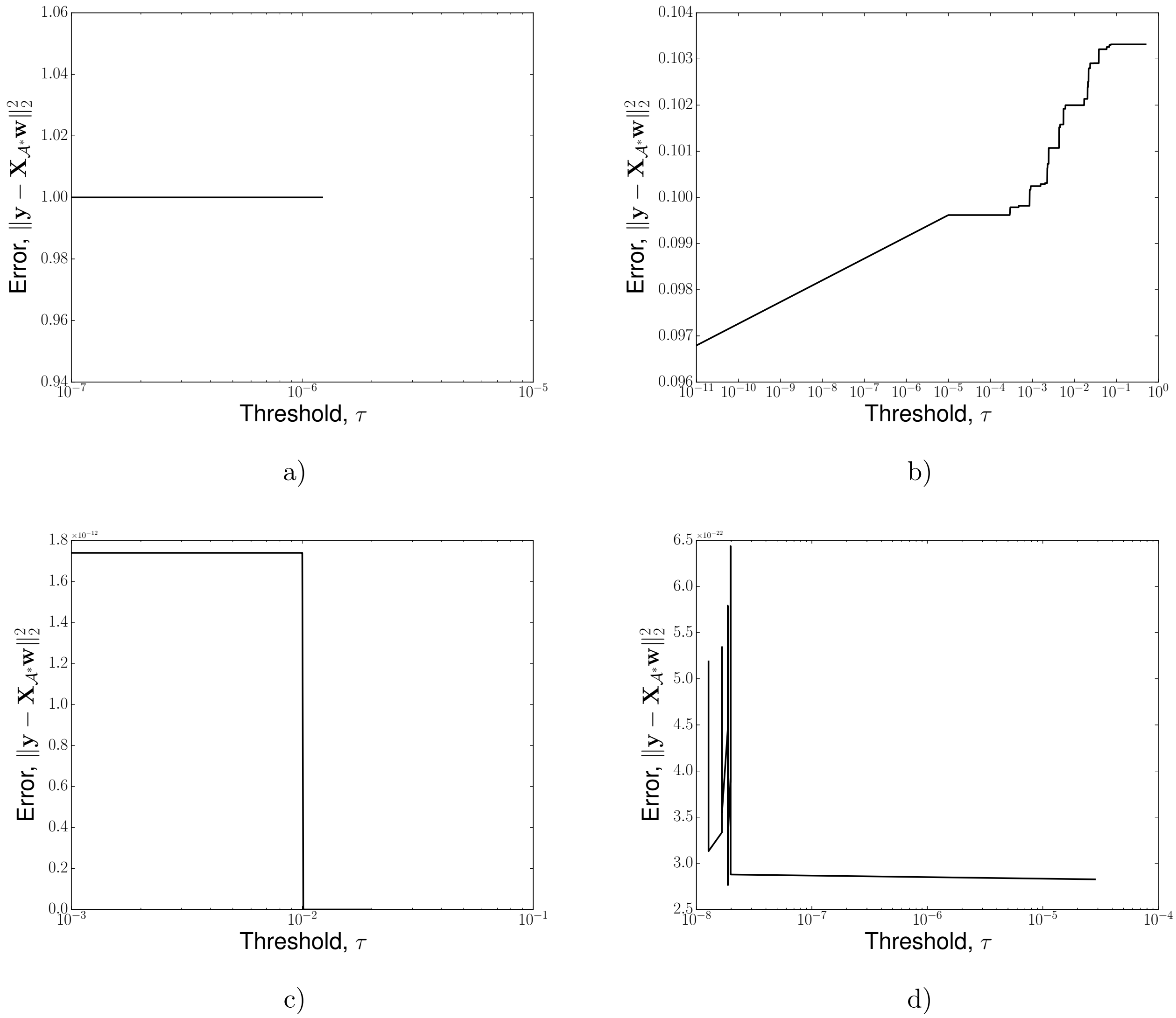


Figure 4: Dependence error function S on the threshold τ for considered types of test data sets: a) inadequate correlated data set, b) adequate random data set, c) adequate redundant data set, d) adequate and correlated data set.

[4] I. Guyon, An introduction to variable and feature selection, Journal of Machine Learning Research 3 (2003) 1157–1182.

[5] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, Journal of the Royal Statistical Society, Series B 67 (2005) 301–320.

[6] M. El-Dereny, N. Rashwan, Solving multicollinearity problem using ridge regression models, International Journal of Contemporary Mathematical Sciences 6 (2011) 585–600.

[7] A. Katrutsa, V. Strijov, Stress test procedure for feature selection algorithms, Chemometrics and Intelligent Laboratory Systems 142 (2015) 172–183.

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

[9] M. A. Hall, Correlation-based feature selection for machine learning, Ph.D. thesis, The University of Waikato (1999).

[10] P. A. Estévez, M. Tesmer, C. A. Perez, J. M. Zurada, Normalized mutual information feature selection, Neural Networks, IEEE Transactions on 20 (2) (2009) 189–201.

[11] M. Grant, S. Boyd, CVX: Matlab software for disciplined convex programming, version 2.1, `http://cvxr.com/cvx` (Mar. 2014).

[12] M. Grant, S. Boyd, Graph implementations for nonsmooth convex programs, in: V. Blondel, S. Boyd, H. Kimura (Eds.), Recent Advances in Learning and Control, Lecture Notes in Control and Information Sciences, Springer-Verlag Limited, 2008, pp. 95–110, `http://stanford.edu/~boyd/graph_dcp.html`.

[13] T. Naghibi, S. Hoffmann, B. Pfister, A semidefinite programming based search strategy for feature selection with mutual information measure, IEEE transactions on pattern analysis and machine intelligence 37 (8) (2015) 1529–1541.

[14] R. G. Askin, Multicollinearity in regression: review and examples, Journal of Forecasting 1 (3) (1982) 281–292.

[15] E. E. Leamer, Multicollinearity: A bayesian interpretation, The Review of Economics and Statistics 55 (3) (1973) 371–80.

[16] D. A. Belsley, E. Kuh, R. E. Welsch, Regression diagnostics: Identifying influential data and sources of collinearity, John Wiley & Sons, New York, 2005.

[17] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, et al., Least angle regression, The Annals of statistics 32 (2) (2004) 407–499.

[18] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society, Series B 58 (1994) 267–288.

[19] F. E. Harrell, Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis, Springer, 2001.

[20] P. Ghamisi, J. A. Benediktsson, Feature selection based on hybridization of genetic algorithm and particle swarm optimization, IEEE Geoscience and Remote Sensing Letters 12 (2) (2015) 309–313.

[21] R. K. Paul, Multicollinearity: Causes, effects and remedies, Tech. rep., Working paper, unknown date. Accessed Apr. 23, 2013, http://pb8.ru/7hy (2006).

[22] S. Paul, S. Das, Simultaneous feature selection and weighting–an evolutionary multi-objective optimization approach, Pattern Recognition Letters 65 (2015) 51–59.

[23] S. G. Gilmour, The interpretation of Mallows's C_p-Statistic, The Statistician (1996) 49–56.

[24] A. D. McQuarrie, C.-L. Tsai, Regression and time series model selection, World Scientific, 1998.

[25] Near infrared spectra of diesel fuels, bp50, `http://www.eigenvector.com/data/SWRI/index.html`.