Katrutsa A.M., Strijov V.V. Stresstest procedure for feature selection algorithms // Chemometrics and Intelligent Laboratory Systems, 2015, 142 : 172-183.

# Stresstest procedures for feature selection algorithms*

A. M. Katrutsa[1,2] and V. V. Strijov[1]

[1] *Moscow Institute of Physics and Technology, Institutskiy lane 9, Dolgoprudny city, 141700, Russian Federation*

[2] *Skolkovo Institute of Science and Technology, Novaya St., 100, Karakorum Building, 4th floor, Skolkovo, 143025, Russian Federation*

**Abstract**

This study investigates the multicollinearity problem and the performance of feature selection methods in case of datasets have multicollinear features. We propose a stresstest procedure for a set of feature selection methods. This procedure generates test data sets with various configurations of the target vector and features. A number of some multicollinear features are inserted in every configuration. A feature selection method results a set of selected features for given test data set. To compare given feature selection methods the procedure uses several quality measures. A criterion of the selected features redundancy is proposed. This criterion estimates number of multicollinear features among the selected ones. To detect multicollinearity it uses the eigensystem of the parameter covariance matrix. In computational experiments we consider the following illustrative methods: Lasso, ElasticNet, LARS, Ridge and Stepwise and determine the best one, which solve the multicollinearity problem for every considered configuration of dataset.

**Keywords:** regression analysis, feature selection methods, multicollinearity, test data sets, the criterion of the selected features redundancy.

## 1 Introduction

This study is devoted to multicollinearity problem and develops a testing procedure for feature selection methods. Assume that data sets have multicollinear features. *Multicollinearity* is a strong correlation between the features, which affect the target vector simultaneously. The multicollinearity reduces the stability of the parameter estimations. The multicollinearity problem, detection methods and methods to solve this problem are discussed in [1, 2, 3]. The

*This publication is based on work funded by Skolkovo Institute of Science and Technology (Skoltech) in the within the framework of the SkolTech/MITInitiative.

parameter vector estimation is called *stable* if a small change of the parameter vector leads to a small change of the target vector estimation.

This study proposes a test procedure for feature selection methods. It uses the various configurations of the target vector and features to construct the test dataset. This procedure is used to compare feature selection methods and to reveal pros and cons of them.

We solve the linear model selection problem. This problem is formulated as the feature selection problem, where selected features fit the target vector in the best way to form the most stable model. The model *stability* is defined as the condition number logarithm of the estimation of the model parameter covariance matrix.

It draws generated test data sets including multicollinear features, features correlated to the target vector, orthogonal features and features orthogonal to the target vector. Setting the cardinality of these feature sets gives opportunity to generate data sets with various features and target vector configuration. This test data sets generation procedure investigates how the considered feature selection method effectiveness depend on continuously increasing the parameter of multicollinearity.

We propose a criterion to rank feature selection methods according to their resistance to multicollinearity. The proposed criterion estimates the number of multicollinear features among the selected ones for given limit value of error function. The feature selection methods are ranked according to increasing the number of multicollinear features in the set of the selected ones. The best method selects features with the minimum number of multicollinear features.

**Related works.** Feature selection methods are used to solve the multicollinearity problem in regression [9]. Also, they are used in the following data mining problems: dimensionality reduction [4, 5], simplification usage of the standard machine learning algorithms [6], removing irrelevant features [7] and increasing the generalisation ability of applying algorithm [8].

The feature selection methods minimize their error functions, that show the quality of the selected subset of features. The papers [10, 11, 12] review existing feature selection methods, classify them according to error functions and optimum feature subset search strategies.

In the presence of multicollinearity in a dataset the feature selection methods improve the parameter estimations stability and reduce their variance. The feature selection methods are based on either regularizators or add-del features strategies. For example, the methods using regularizators are ridge regression [13], where the regularizator is the weighted euclidean norm of the parameter vector, Lasso [14] and LARS [15], where the regularizator is the weighted sum of the vector parameter elements, Elastic net [16], where the regularizator is the linear combination of the two previous regularizators. Stepwise uses the F-test to detect the most significant feature and add it or the least significant to remove it.

The most related paper about feature selection test [9] proposes a data sets generation procedure and quality measure to evaluate the feature selection method quality. However it doesn't evaluate the quality measure while the parameter of multicollinearity and data set

parameters are changing continuously.

## 2 Feature selection problem statement

Let $\mathfrak{D} = \{(\mathbf{X}, \mathbf{y})\}$ be the given data set, where the design matrix

$$\mathbf{X} = [\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_j, \ldots, \boldsymbol{\chi}_n],\ \mathbf{X} \in \mathbb{R}^{m \times n} \text{ and } j \in \mathcal{J} = \{1, \ldots, n\}.$$

The vector $\boldsymbol{\chi}_j$ is called the $j$-th feature and the vector $\mathbf{y} = [y_1, \ldots, y_m]^\intercal \in \mathbb{Y} \subset \mathbb{R}^m$ is called the target vector. Assume that the target vector $\mathbf{y}$ and design matrix $\mathbf{X}$ are related through the following equation:

$$\mathbf{y} = \mathbf{f}(\mathbf{w}, \mathbf{X}) + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{f}$ maps the cartesian product of the feasible parameter space and the space of the $m \times n$ matrices to the target vector domain, and $\boldsymbol{\varepsilon}$ is the residual vector. The data fit problem is to estimate the parameter vector $\mathbf{w}^*$,

$$\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathbb{R}^n} S(\mathbf{w} | \mathfrak{D}_{\mathcal{L}}, \mathcal{A}, \mathbf{f}), \tag{2}$$

where $S$ is the error function. The set $\mathfrak{D}_{\mathcal{L}} \subset \mathfrak{D}$ is a training set and the set $\mathcal{A} \subseteq \mathcal{J}$ is the *active index set* used in computing the error function $S$. In the stresstest procedure we use the quadratic error function

$$S = \|\mathbf{y} - \mathbf{f}(\mathbf{w}, \mathbf{X})\|_2^2 \tag{3}$$

and the linear regression function $\mathbf{f}(\mathbf{w}, \mathbf{X}) = \mathbf{X}\mathbf{w}$. The introduced stresstest procedure could be applied to the generalised linear model selection algorithms, where the model is $\mathbf{f} = \boldsymbol{\mu}^{-1}(\mathbf{X}\mathbf{w})$ and $\boldsymbol{\mu}$ is a link function.

**Definition 2.1** Let $\mathcal{A}^*$ denote *the optimum index set*, the solution of the problem

$$\mathcal{A}^* = \arg\min_{\mathcal{A} \subseteq \mathcal{J}} S_{\mathfrak{m}}(\mathcal{A} | \mathbf{w}^*, \mathfrak{D}_{\mathcal{C}}, \mathbf{f}), \tag{4}$$

where $\mathfrak{D}_{\mathcal{C}} \subset \mathfrak{D}$ is the test set, $\mathbf{w}^*$ is the solution of the problem (2) and $S_{\mathfrak{m}}$ is an error function corresponding to a feature selection method $\mathfrak{m}$ (5).

The feature selection problem (4) is to find the optimum index set $\mathcal{A}^*$. It must exclude indices of noisy and multicollinear features. It is expected that if one uses features indexed by the set $\mathcal{A}^*$ then it brings more stable solution of the problem (2), in comparison to the case of $\mathcal{A} \equiv \mathcal{J}$.

In the computational experiment we consider the feature selection methods from the set $\mathfrak{M} = \{\text{Lasso, LARS, Stepwise, ElasticNet, Ridge}\}$.

**Definition 2.2** A feature selection method $\mathfrak{m} \in \mathfrak{M}$ is a map from the complete index set $\mathcal{J}$ to active index set $\mathcal{A} \subseteq \mathcal{J}$:

$$\mathfrak{m} : \mathcal{J} \to \mathcal{A}. \tag{5}$$

According to this definition we consider the terms feature selection problem and the model selection problem to be synonyms.

**Definition 2.3** Let a model be a pair $(\mathbf{f}, \mathcal{A})$, where $\mathcal{A} \subseteq \mathcal{J}$ is an index set. The model selection problem is to find the optimum pair $(\mathbf{f}^*, \mathcal{A}^*)$ which minimizes the error function $S$ (3).

**Definition 2.4** Call *the model complexity* $C$ the cardinality of the active index set $\mathcal{A}$, number of the selected features:

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

**Definition 2.5** Define *the model stability* $R$ be logarithm of the condition number $\kappa$ of the matrix $\mathbf{X}^\intercal\mathbf{X}$:

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

where $\lambda_{\max}$ and $\lambda_{\min}$ are the maximum and the minimum non-zero eigenvalue of the matrix $\mathbf{X}^\intercal\mathbf{X}$. The features with indices from the corresponding active set $\mathcal{A}$ are used in computing the condition number $\kappa$.

# 3 Multicollinearity analysis in feature selection

In this section we give definitions of multicollinear features, correlated features and features correlated with the target vector. In the following subsections we list and study the multicollinearity criteria.

Assume that the features $\boldsymbol{\chi}_j$ and the 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 3.1** 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

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

The smaller $\delta$ the higher *degree of multicollinearity.*

**Definition 3.2** Call 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 $\delta_{ij} = \delta_{ji}$. In the special case $a_k = 0$ $k \neq j$ and $a_k = 1$ $k = j$ the inequalities (8) and (7) are identically.

**Definition 3.3** A feature $\boldsymbol{\chi}_j$ is called *correlated with the target vector* $\mathbf{y}$ if there exists sufficiently small positive number $\delta_{yj} > 0$ such that

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

Further used the following notations RSS (Residual Sum of Squares) and TSS (Total Sum of Squares):

$$\text{RSS} = S(\mathfrak{D}_{\mathcal{L}}, \mathbf{w}^*) = \|\boldsymbol{\varepsilon}\|_2^2 \quad \text{and} \quad \text{TSS} = \sum_{i=1}^{m}(y_i - \overline{y})^2, \text{ where } \overline{y} = \frac{1}{m}\sum_{i=1}^{m} y_i. \tag{9}$$

## 3.1 Variance inflation factor

The variance inflation factor $\text{VIF}_j$ is used as a multicollinearity indicator [17]. The $\text{VIF}_j$ is defined for $j$-th feature and shows a linear dependence between $j$-th feature and the other features.

To compute $\text{VIF}_j$ 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 $\mathcal{J} = \mathcal{J} \setminus j$. The functions RSS and TSS are computed similar to (9). The $\text{VIF}_j$ is computed with the following equation:

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

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

According to [17] any $\text{VIF}_j \gtrsim 5$ indicates that the associated elements of the vector $\mathbf{w}^*$ are poorly estimated because of multicollinearity. Denote by VIF the maximum value of $\text{VIF}_j$ for all $j \in \mathcal{J}$:

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

However, $\text{VIF}_j$ can be infinitely large for some features. In this case it is impossible to determine which features must be removed from the active set. This is major disadvantage of the variance inflation factor.

Another multicollinearity indicator is the condition number $\kappa$ of the matrix $\mathbf{X}^\intercal\mathbf{X}$. The condition number is defined as:

$$\kappa = \frac{\lambda_{\max}}{\lambda_{\min}},$$

where the $\lambda_{\max}$ and $\lambda_{\min}$ are the maximum and minimum non-zero eigenvalues of the matrix $\mathbf{X}^\intercal\mathbf{X}$.

The condition number shows how much does the matrix $\mathbf{X}^\intercal\mathbf{X}$ close to the singular matrix. The larger $\kappa$ the more ill-conditioned matrix $\mathbf{X}^\intercal\mathbf{X}$.

## 3.2 The Belsley criterion

To detect and remove indices of the multicollinear features from the active index set we state the direct optimization problem using the Belsley criterion. We propose the new criterion

to compare feature selection methods: *the criterion of the selected features redundancy.* This criterion uses the maximum cardinality of the redundant index set, which can be removed within the error function does not raised above given value. The features are removed according to the Belsley criterion described below. The formal definition of the the maximum cardinality of the redundant index set is given by (16).

Assume that the parameter vector $\mathbf{w} \in \mathbb{R}^n$ has the multivariate normal distribution with the expectation $\mathbf{w}_{\mathrm{ML}}$ and the covariance matrix $\mathbf{A}^{-1}$,

$$\mathbf{w} \sim \mathcal{N}(\mathbf{w}_{\mathrm{ML}}, \mathbf{A}^{-1}).$$

The estimation $\hat{\mathbf{A}}^{-1}$ of the covariance matrix $\mathbf{A}^{-1}$ in the linear model is

$$\hat{\mathbf{A}}^{-1} = (\mathbf{X}^\intercal \mathbf{X})^{-1}.$$

To inverse $\mathbf{X}^\intercal \mathbf{X}$ we use the singular value decomposition of the $m \times n$ matrix $\mathbf{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\intercal$, where $\mathbf{U}$ and $\mathbf{V}$ are the orthogonal matrices, and $\mathbf{\Lambda}$ is the diagonal matrix with the singular values $\sqrt{\lambda_i}$ on the diagonal, such that

$$\sqrt{\lambda_1} \geq \ldots \geq \sqrt{\lambda_i} \geq \ldots \geq \sqrt{\lambda_r} > 0,$$

where $i = 1, \ldots, r$ and $r = \min(m, n)$. Thus, the inversion $\mathbf{X}^\intercal \mathbf{X}$ is following:

$$(\mathbf{X}^\intercal \mathbf{X})^{-1} = \mathbf{V}\mathbf{\Lambda}^{-2}\mathbf{V}^{-1}.$$

The columns of the matrix $\mathbf{V}$ is the eigenvectors and the squares of the singular values $\lambda_i$ are the eigenvalues of the matrix $\mathbf{X}^\intercal \mathbf{X}$ since $\mathbf{X}^\intercal \mathbf{X} = \mathbf{V}\mathbf{\Lambda}^\intercal \mathbf{U}^\intercal \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\intercal = \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^\intercal$ and $\mathbf{X}^\intercal \mathbf{X}\mathbf{V} = \mathbf{V}\mathbf{\Lambda}^2$.

**Definition 3.4** The ratio of the maximum eigenvalue $\lambda_{\max}$ to the $i$-th eigenvalue $\lambda_i$ is called *the condition index* $\eta_i$

$$\eta_i = \frac{\lambda_{\max}}{\lambda_i}.$$

The large value of $\eta_i$ indicates the close-to-linear relation between the features. The larger value of $\eta_i$ the closer relation between features to linear.

The variance of the vector $\mathbf{w}^*$ elements are estimated as diagonal entries of the matrix $\mathbf{X}^\intercal \mathbf{X} = \mathbf{V}\mathbf{\Lambda}^2\mathbf{V}^\intercal$:

$$\mathrm{Var}(w_i) = \sum_{j=1}^{n} \frac{v_{ij}^2}{\lambda_j^2}.$$

**Definition 3.5** *The coefficient variance proportion* $q_{ij}$ is the $j$-th feature contribution to the variance of the $i$-th element of the optimal parameter vector $\mathbf{w}^*$. The formal definition of the coefficient variance proportion $q_{ij}$ is

$$q_{ij} = \frac{v_{ij}^2/\lambda_j^2}{\sum\limits_{j=1}^{n} v_{ij}^2/\lambda_j^2},$$

where $[v_{ij}] = \mathbf{V}$ and $\lambda_j$ is the eigenvalue of the matrix $\mathbf{X}^\intercal \mathbf{X}$.

By this definition, larger value of the coefficient variance proportion shows close-to-linear relation of the features.

To find the feature $\boldsymbol{\chi}_{j^*}$ from the multicollinear features set the Belsley criterion uses both the condition index and the coefficient variance proportion in the following way. One must find the feature index $i^*$ such that

$$i^* = \arg\max_{i \in \mathcal{A}^*} \eta_i,$$

and using this index $i^*$ find another feature index $j^*$ such that

$$j^* = \arg\max_{j \in \mathcal{A}^*} q_{i^* j}. \tag{10}$$

According to the Belsley criterion the $j^*$-th feature gives the largest contribution to variance of the $i^*$-th element of the parameter vector $\mathbf{w}^*$. There is the close-to-linear relation between the $j^*$-th feature and the other ones. Therefore, we find the multicollinear feature which has to be removed from the active index set $\mathcal{A}^*$.

# 4 Test data set generation procedure

To test feature selection methods we propose a procedure to construct test synthetic data sets. To define a test data set let $\mathcal{P}_f, \mathcal{P}_y, \mathcal{C}_f, \mathcal{C}_y, \mathcal{R}$ be the following index sets:

1) the index set $\mathcal{P}_f$ labels the orthogonal features;

2) the index set $\mathcal{P}_y$ labels the features which are orthogonal to the target vector $\mathbf{y}$;

3) the index set $\mathcal{C}_f$ labels the multicollinear features;

4) the index set $\mathcal{C}_y$ labels the correlated with the target vector $\mathbf{y}$ features;

5) the index set $\mathcal{R}$ labels the random features.

Denote the cardinalities of the declared index sets by

$$|\mathcal{P}_f| = p_f, \qquad |\mathcal{P}_y| = p_y, \qquad |\mathcal{C}_f| = c_f, \qquad |\mathcal{C}_y| = c_y, \qquad |\mathcal{R}| = r.$$

To control the *degree of multicollinearity* (7) we use the parameter of multicollinearity $k$. If $k = 1$, the features are multicollinear and the degree of multicollinearity is high. If $k = 0$, they are orthogonal and the degree of multicollinearity is low.

The basic configurations of the test data sets are considered below:

1) an inadequate and correlated data set;

2) an adequate and random data set;

3) an adequate and redundant data set;

4) an adequate and correlated data set.

1. The first basic configuration of the test data set consists of the target vector $\mathbf{y}$ and features $\boldsymbol{\chi}_j$ with their indices $j$ from both index set $\mathcal{C}_f$ of the multicollinear features and index set $\mathcal{P}_y$ of the orthogonal to the target vector $\mathbf{y}$ features:

$$\left\| \boldsymbol{\chi}_i - \sum_{l \in \mathcal{A}} \alpha_l \boldsymbol{\chi}_l \right\|_2^2 < \delta, \quad i \in \mathcal{J}, \quad i \notin \mathcal{A} \subset \mathcal{J}, \quad \langle \mathbf{y}, \boldsymbol{\chi}_j \rangle = 0, \quad j \in \mathcal{J}, \quad \mathcal{J} = \mathcal{P}_y \cap \mathcal{C}_f. \tag{11}$$

The fig. 1 shows a configuration of this dataset. It is called *the inadequate and correlated data set*.

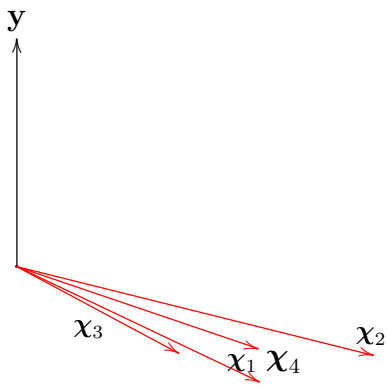



Figure 1: The inadequate and correlated data set

2. The second basic configuration of the test data set consists of the target vector $\mathbf{y}$ and the features $\boldsymbol{\chi}_j$ which are generated from the multivariate uniform distribution on the $r$-dimensional unit hypercube and one of these features $\boldsymbol{\chi}_i$ correlates with the target vector $\mathbf{y}$:

$$\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_j, \ldots, \boldsymbol{\chi}_r \sim \mathcal{U}[0,1]^m, \quad \|\mathbf{y} - \boldsymbol{\chi}_i\|_2^2 < \delta, \quad j \in \mathcal{J} = \mathcal{R}, \quad |\mathcal{R}| = r. \tag{12}$$

The fig. 2 shows a configuration of this dataset. It is called *the adequate and random data set*.

3. The third basic configuration of the test data set consists of the target vector $\mathbf{y}$ and the features $\boldsymbol{\chi}_j$ correlate with each other and approximate the target vector $\mathbf{y}$:

$$\left\|\boldsymbol{\chi}_i - \boldsymbol{\chi}_j\right\|_2^2 < \delta_{ij}, \quad i,j \in \mathcal{J}, \quad \|\mathbf{y} - \boldsymbol{\chi}_j\|_2^2 < \delta, \quad j \in \mathcal{J}, \quad \mathcal{J} = \mathcal{C}_y \tag{13}$$

The fig. 3 shows a configuration of this data set. It is called *the adequate and redundant data set*.

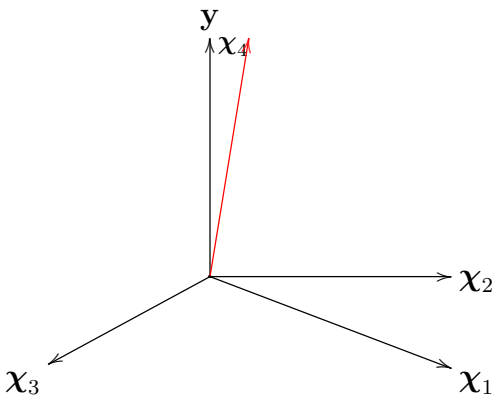



Figure 2: The adequate and random data set

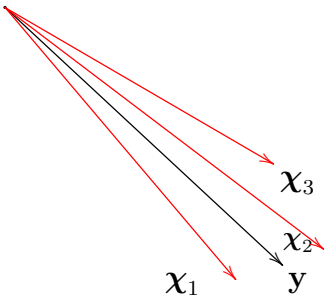



Figure 3: The adequate and redundant data set

4. The forth basic configuration of the test data set consists of the target vector $\mathbf{y}$ and the features with indices from the set $\mathcal{J}$. The index set $\mathcal{J}$ is the union of the index set $\mathcal{P}_f$ of the orthogonal features and the index set $\mathcal{C}_f$ of the multicollinear features. Particularly, the set $\mathcal{C}_f$ contains the indices of the features, which are correlated with some of the orthogonal features. At the same time, the target vector $\mathbf{y}$ equals the linear combination of the features $\boldsymbol{\chi}_j,\ j \in \mathcal{P}_f$:

$$\langle \boldsymbol{\chi}_i, \boldsymbol{\chi}_j \rangle = 0, \quad i,j \in \mathcal{P}_f, \quad \|\boldsymbol{\chi}_i - \boldsymbol{\chi}_j\|_2^2 < \delta_{ij}, \quad i \in \mathcal{P}_f,\ j \in \mathcal{C}_f, \quad \mathbf{y} = \sum_{j \in \mathcal{P}_f} a_j \boldsymbol{\chi}_j, \tag{14}$$
$$\mathcal{J} = \mathcal{P}_f \cup \mathcal{C}_f.$$

   The fig. 4 shows a configuration of this data set. It is called *the adequate and correlated data set*.

You can obtain any other test data sets by the one or combination of the following ways: change the cardinality of the index sets described above, change the parameter of multicollinearity $k$, combine the basic configuration within the one test data set.

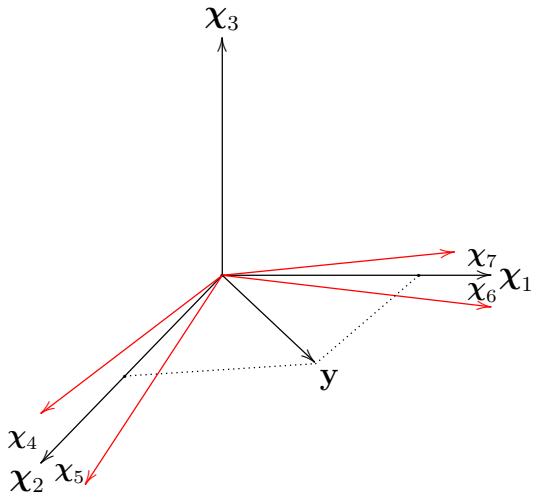



Figure 4: The adequate and correlated data set

# 5 The criterion of the selected features redundancy

To compare the feature selection methods we propose *the criterion of the selected features redundancy*. It estimates the number of the multicollinear feature indices in the active set. Denote by $s_0$ a limit value of the error function $S$. By definition 2.2 the feature selection method returns the feature indices subset $\mathcal{A} \subset \mathcal{J}$. Estimate the optimum parameter vector $\mathbf{w}^*_{\mathcal{A}}$ with features indexed by the set $\mathcal{A}$. Denote by $h$ the maximum cardinality of the index set $\mathcal{J}_h \subseteq \mathcal{A}$ such that the value of the error function $S$ is less or equal to $s_0$

$$S(\mathcal{J}_h|\mathbf{w}^*_h, \mathfrak{D}) \leq s_0,$$

where $\mathbf{w}^*_h$ is a vector, such that the $i$-th element of $\mathbf{w}^*_h$ equals the $i$-th element of $\mathbf{w}^*_{\mathcal{A}^*}$, if $i \in \mathcal{J}_h$ and zero, otherwise. The maximum cardinality $h$ is the solution of the problem

$$h = \underset{S(\mathcal{J}_h|\mathbf{w}^*_h, \mathfrak{D}) \leq s_0}{\arg\max} |\mathcal{J}_h|. \tag{15}$$

**Definition 5.1** Denote by $d$ the maximum cardinality of the index set of the redundant features, number of the redundant features:

$$d = |\mathcal{A}| - h. \tag{16}$$

The redundant feature indices are found according to the Belsley criterion discussed in the subsection 3.2 as the solution of the problem (10). Remove the obtained feature indices sequentially until the error $S$ is less or equal $s_0$.

*The criterion of the selected features redundancy* ranks the feature selection methods in the following way: the feature selection method $\mathfrak{m}_i$ is better than the feature selection method $\mathfrak{m}_j$ if and only if the corresponding value of $d_i$ is smaller than the corresponding value of $d_j$:

$$d_i < d_j \Leftrightarrow \mathfrak{m}_i \succ \mathfrak{m}_j.$$

**Standard criteria to compare linear regression models.** Previously in [17, 18] the authors propose the following criteria to compare linear regression models.

1. The adjusted coefficient of determination $R^2_{\text{adj}}$ considers adding redundant features and defined as
$$R^2_{\text{adj}} = 1 - \frac{\text{RSS}/(m-k)}{\text{TSS}/(m-1)}.$$
The closer $R^2_{\text{adj}}$ to one the better model fits the target vector.

2. The Mallow's $C_p$ criterion trades off the RSS and the number of features $p$. The Mallow's $C_p$ defined as
$$C_p = \frac{\text{RSS}_p}{\text{RSS}} - m + 2p,$$
where $\text{RSS}_p$ is similar to RSS, but computed with $p$ features only. In terms of this criterion the smaller $C_p$ the better feature subset.

3. Information criterion BIC defined as
$$\text{BIC} = \text{RSS} + p \log m.$$
The smaller value of BIC the better model fits the target vector.

# 6 Computational experiment

We execute the computational experiment in four stages. The first stage compares illustrative feature selection methods on various quality measures. The limit error function $s_0 = 0.5$ and the parameter of multicollinearity $k = 0.2$ and $k = 0.8$. The second stage performs VIF-analysis of the multicollinearity. This analysis is illustrated by the plots of the parameter of multicollinearity $k$ versus VIF. These plots are obtained for the inadequate and correlated, adequate and redundant and adequate and correlated data sets. A plot for an adequate and random data set was omitted because there is no multicollinearity. The third stage investigates the number of the redundant features $d$ for given limit error function values $s_0$. The obtained pairs $d$ and $s_0$ are shown for the inadequate and correlated, adequate and redundant, adequate and correlated data sets. The parameter of multicollinearity $k = 0.2$ and $k = 0.8$. The fourth stage compares illustrative feature selection methods on complexity $C$ and stability $R$ of the models, which are obtained using them. For every kind of data set we find the best one feature selection method that gives the simplest model with the highest stability.

To carry out the experiments we generate four test data sets according to the equations (11), (12), (13) and (14). The features and the target vector are normalized (6) before the experiments. The elements of the optimum vector $\mathbf{w}^*$ less $10^{-6}$ are assumed insignificant and equal zero. The size $m$ and $n$ of the design matrix $\mathbf{X}$ is listed below for every generated data set:

1) inadequate and correlated data sets: $n = p_y = 50$, $m = 1000$;

2) adequate and random data sets: $n = r = 50$, $m = 1000$;

3) adequate and redundant data sets: $n = c_y = 50$, $m = 1000$;

4) adequate and correlated data sets: $p_f = 10$, $c_f = 40$, $m = 1000$.

We compare the following feature selection methods: LARS, Lasso, ElasticNet, Ridge and Stepwise. For the ElasticNet penalty we use the weight 0.5 both in Lasso and Ridge penalty. The dash in a table row means that corresponding feature selection method doesn't select any feature.

## 6.1 Comparing feature selection methods

This stage compares feature selection methods on the various quality measures. The limit error function $s_0 = 0.5$ and the parameter of multicollinearity $k = 0.2$ and $k = 0.8$. The results of comparisons are in the tables 1, 2, 3 and 4. The feature selection methods are sorted in the tables according to simultaneous increasing of the maximum number of the redundant features $d$ and RSS.

Table 1: Quality measures for the inadequate and correlated data sets

| Method | $d$ | $C_p$ | RSS | $\kappa$ | VIF | $R^2_{\text{adj}}$ | BIC |
|---|---|---|---|---|---|---|---|
| $k = 0.2$ | | | | | | | |
| Lasso | 0 | −997 | 1 | 3.84 | 1.05 | −3.32 | 314.62 |
| Ridge | 0 | −997 | 1 | 4.13 | 1.05 | −3.31 | 346.39 |
| LARS | 0 | −997 | — | — | — | — | — |
| Stepwise | 0 | −997 | 1 | 4.13 | 1.05 | −3.41 | 346.41 |
| Elastic Net | 0 | −997 | 1 | 3.84 | 1.05 | −3.32 | 314.32 |
| $k = 0.8$ | | | | | | | |
| Lasso | 0 | −997 | 1 | 717.8 | 16.6 | −3.32 | 310.48 |
| Ridge | 0 | −997 | 1 | 801 | 16.6 | −3.31 | 346.39 |
| LARS | — | −997 | — | — | — | — | — |
| Stepwise | 0 | −997 | 1.68 | 801 | 16.6 | −6.22 | 347.01 |
| Elastic Net | 0 | −997 | 1 | 717.8 | 16.6 | −3.32 | 310.48 |

## 6.2 VIF-analysis of multicollinearity

This stage obtains the dependence of VIF on the parameter of multicollinearity $k$ for all considered data sets assuming $\mathcal{J} = \mathcal{A}$ and a number of illustrative feature selection methods.

The fig. 5 shows VIF-analysis of multicollinearity for the inadequate and correlated data sets and all illustrative feature selection methods. None of them solve the multicollinearity

Table 2: Quality measures for the adequate and random data sets

| Method | $d$ | $C_p$ | RSS | $\kappa$ | VIF | $R^2_{\text{adj}}$ | BIC |
|---|---|---|---|---|---|---|---|
| Lasso | 0 | $7 \cdot 10^6$ | $8.50 \cdot 10^{-4}$ | 1 | 0.25 | 1 | 6.9 |
| Elastic Net | 0 | $8.76 \cdot 10^{-4}$ | $8.76 \cdot 10^{-4}$ | 1 | 0.25 | 1 | 6.9 |
| Ridge | 0 | $7.97 \cdot 10^9$ | 0.97 | 1 | 0.25 | $-3$ | 7.88 |
| LARS | 0.2 | $-997$ | $1.3 \cdot 10^{-10}$ | 2.19 | 0.32 | 1 | 8.29 |
| Stepwise | 4.6 | $-997$ | $1.33 \cdot 10^{-10}$ | 28.86 | 0.89 | 1 | 53.88 |

Table 3: Quality measures for the adequate and redundant data sets

| Method | $d$ | $C_p$ | RSS | $\kappa$, $\cdot 10^8$ | VIF, $\cdot 10^7$ | $R^2_{\text{adj}}$ | BIC |
|---|---|---|---|---|---|---|---|
| $k = 0.2$ | | | | | | | |
| Ridge | 0 | $2.3 \cdot 10^9$ | 0.97 | 24 | 1.14 | $-3.17$ | 346.36 |
| Lasso | 1 | $2 \cdot 10^6$ | $8.5 \cdot 10^{-4}$ | 0.95 | 0.58 | 1 | 13.82 |
| Elastic Net | 3.2 | $2 \cdot 10^6$ | $8.5 \cdot 10^{-4}$ | 2.8 | 0.97 | 1 | 41.45 |
| Stepwise | 36 | $-997$ | $4.22 \cdot 10^{-10}$ | 24 | 1.14 | 1 | 345.39 |
| LARS | 36 | $-997$ | $4.22 \cdot 10^{-10}$ | 24 | 1.14 | 1 | 345.39 |
| $k = 0.8$ | | | | | | | |
| Lasso | 0 | $5.16 \cdot 10^8$ | $8.5 \cdot 10^{-4}$ | 1 | 0.24 | 1 | 6.9 |
| Ridge | 0 | $5.9 \cdot 10^{11}$ | 0.97 | $6.07 \cdot 10^{11}$ | $2.9 \cdot 10^9$ | $-3.17$ | 346.36 |
| Elastic Net | 3.2 | $5.16 \cdot 10^8$ | $8.5 \cdot 10^{-4}$ | $7.3 \cdot 10^{10}$ | $2.5 \cdot 10^9$ | 1 | 41.45 |
| Stepwise | 36 | $-997$ | $1.73 \cdot 10^{-12}$ | $6.07 \cdot 10^{11}$ | $2.9 \cdot 10^9$ | 1 | 345.39 |
| LARS | 36 | $-997$ | $1.65 \cdot 10^{-12}$ | $6.07 \cdot 10^{11}$ | $2.9 \cdot 10^9$ | 1 | 345.39 |

Table 4: Quality measures for the adequate and correlated data sets

| Method | $d$ | $C_p$ | RSS | $\kappa$ | VIF | $R^2_{\text{adj}}$ | BIC |
|---|---|---|---|---|---|---|---|
| $k = 0.2$ | | | | | | | |
| Stepwise | 1 | $-868.95$ | $5.45 \cdot 10^{-29}$ | 1 | 0.63 | 1 | 13.82 |
| Ridge | 0 | $6 \cdot 10^{30}$ | 0.95 | $8.42 \cdot 10^{15}$ | $1.15 \cdot 10^{23}$ | $-3$ | 210.95 |
| LARS | 1.8 | $5.38 \cdot 10^{29}$ | 0.38 | $2.1 \cdot 10^{16}$ | $3.3 \cdot 10^{30}$ | $-0.62$ | 102.62 |
| Lasso | 18 | $5.84 \cdot 10^{27}$ | $9.18 \cdot 10^{-4}$ | $1.4 \cdot 10^{16}$ | $5.32 \cdot 10^{20}$ | 1 | 150.6 |
| Elastic Net | 17.6 | $5.84 \cdot 10^{27}$ | $9.18 \cdot 10^{-4}$ | $1.4 \cdot 10^{16}$ | $5.32 \cdot 10^{20}$ | 1 | 150.59 |
| $k = 0.8$ | | | | | | | |
| Stepwise | 1 | $9.4 \cdot 10^5$ | $8.8 \cdot 10^{-25}$ | 1 | 0.63 | 1 | 13.82 |
| Ridge | 0 | $1.8 \cdot 10^{30}$ | 0.95 | $10^{16}$ | $8.65 \cdot 10^{16}$ | $-2.97$ | 152.92 |
| LARS | 1.2 | $10^{30}$ | 0.38 | $3 \cdot 10^{29}$ | $10^{20}$ | $-0.57$ | 108.15 |
| Lasso | 14.8 | $1.73 \cdot 10^{27}$ | $9.2 \cdot 10^{-4}$ | $9.92 \cdot 10^{15}$ | $10^{17}$ | 1 | 150.59 |
| Elastic Net | 15.2 | $1.7 \cdot 10^{27}$ | $9.2 \cdot 10^{-4}$ | $9.92 \cdot 10^{15}$ | $10^{17}$ | 1 | 150.59 |

problem in case of features are correlated and orthogonal to the target vector. Only LARS diagnoses the absence of the relevant features to fit the target vector and returns the empty index set. This is observed in the table 1.

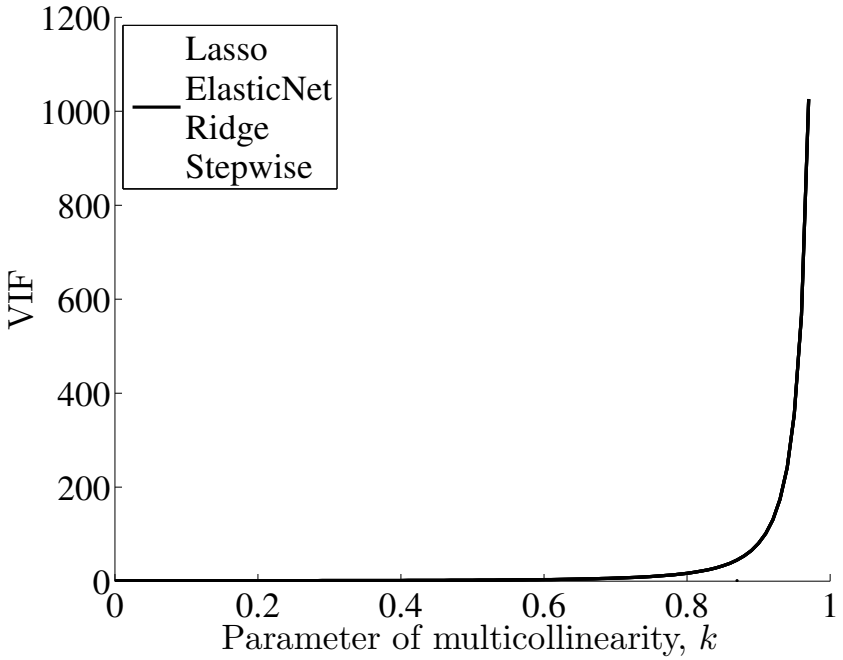



Figure 5: VIF-analysis of multicollinearity for the inadequate and correlated data sets. The increase parameter of multicollinearity $k$ means increase of the degree of multicollinearity.

The fig. 6 shows VIF-analysis of multicollinearity for the adequate and redundant data sets. All methods gives the same VIF curves except Lasso. Using Lasso we observe the sharp decline of the VIF curve since the parameter of multicollinearity $k$ is more than 0.4. It means that there is no linear relation between the features selected by Lasso in the data sets with the parameter of multicollinearity $k \gtrsim 0.4$.

The fig. 7 shows VIF-analysis of multicollinerity for the adequate and correlated data sets. LARS gives strong jumps of the VIF curve (fig. 7 (a)). Therefore, fig. 7 (b) shows the VIF curves for Lasso and ElasticNet. They give the VIF curves similar to LARS, but the jumps of these curves have smaller amplitude. Thus, fig. 7 (c) and 7 (d) show the VIF curves for Stepwise and Ridge. After applying Stepwise method to the adequate and correlated data sets, corresponding VIF $\lesssim 2$ while increasing the parameter of multicollinearity $k$. Hence, Stepwise returns the index set without any linear relations among the corresponding features.

## 6.3 Analysis of the feature selection methods redundancy

The third stage studies the feature selection methods redundancy with the proposed criterion (15) for the previously described data sets. The graphs of the limit error function $s_0$ versus the maximum number of the redundant features $d$ are shown on the fig. 8, 9, 10.

The fig. 8 shows the plots of the limit error function $s_0$ versus the number of the redundant

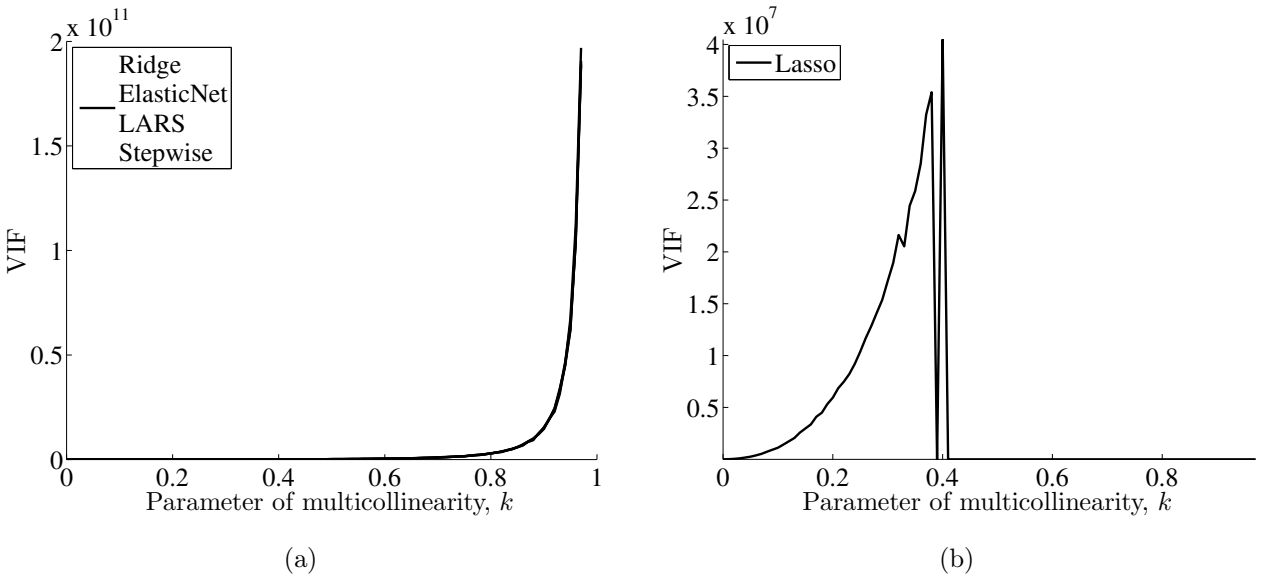



Figure 6: VIF-analysis of multicollinearity for the adequate and redundant data sets used: (a) all considered feature selection methods except Lasso, (b) Lasso only.

features $d$ for the inadequate and correlated data set and parameter of multicollinearity $k = 0.2$ and $k = 0.8$. Since the limit error function $s_0 < 1$, the value of $d = 0$, because of the target vector is orthogonal to the features. Since $s_0 \gtrsim 1$, the value of $d$ increases sharp, because of the limit error function $s_0$ is enough great to remove most features.

The fig. 9 shows the plots of the limit error function $s_0$ versus the maximum number of the redundant features $d$ for the adequate and redundant data set and the parameter of multicollinearity $k = 0.2$ and $k = 0.8$. Lasso selects one feature for $k = 0.2$ and two features for $k = 0.8$ which fit the target vector best way. Therefore the number of the redundant features $d$ equals zero for $k = 0.2$ and one for $k = 0.8$. ElasticNet selects the greater number of the redundant features than Lasso. Ridge shows the stable zero line since $s_0 < 1$ and sharp increasing since $s_0 \gtrsim 1$. This result is similar to the result for the inadequate and correlated data set by the same reason: first $s_0$ is enough small to remove even one feature, later $s_0$ is too great to remove most features. LARS and Stepwise show the smooth increasing of the $d$ while $s_0 < 1$ and constant level around 48 while $s_0 \geq 1$.

The fig. 10 shows plots of the limit error function $s_0$ versus the maximum number of the redundant features $d$ for the adequate and correlated data set and the parameter of multicollinearity $k = 0.2$ and $k = 0.8$. Stepwise gives one redundant feature for all values of the limit error function $s_0$. LARS gives no more five redundant features since increasing the limit error function $s_0$. Lasso and ElasticNet show the increasing of $d$ since $s_0 < 1$ and $d \simeq 20$ since $s_0 \geq 1$. Ridge shows the plot of the $s_0$ versus $d$ similar to the previous data sets, but the value of $d$ starts oscillating since $s_0 \geq 1$.

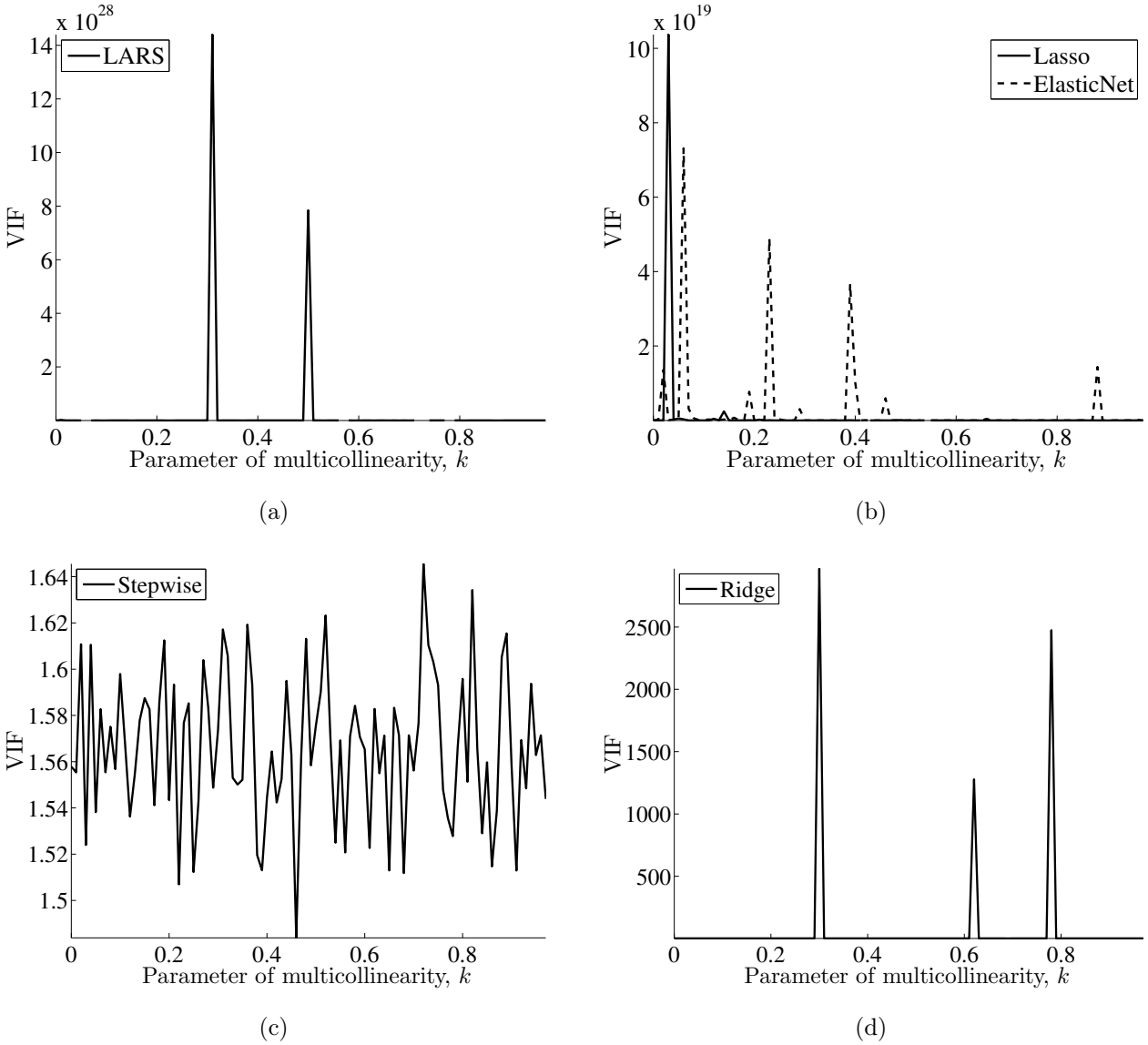


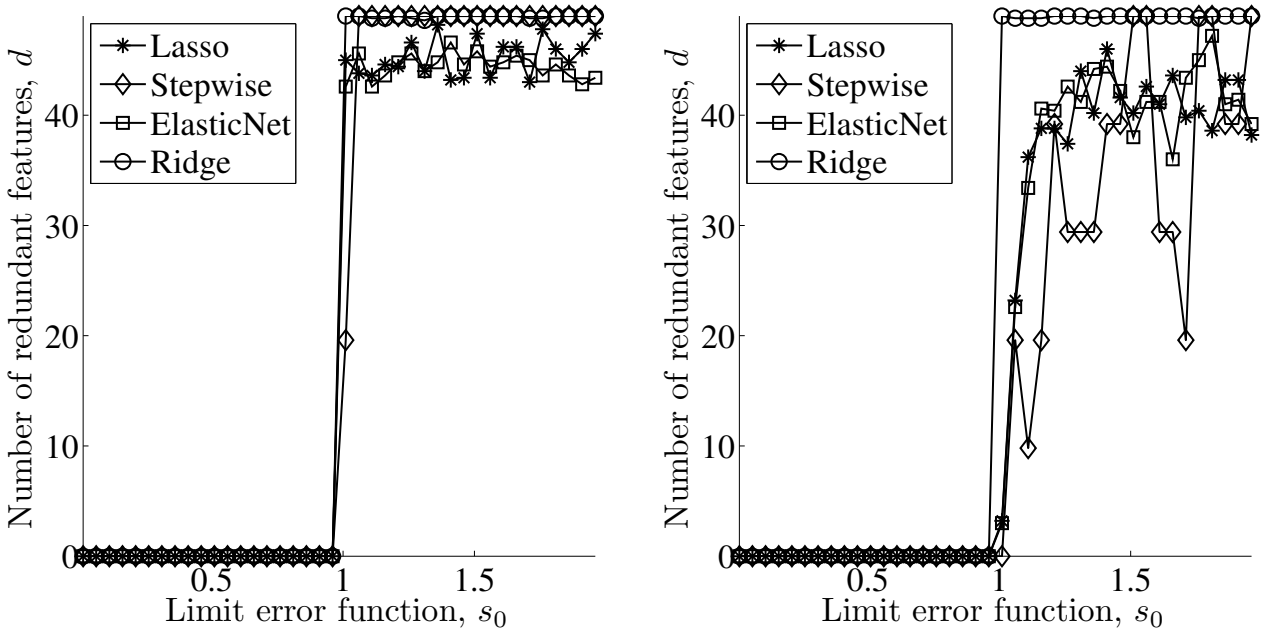

Figure 7: VIF-analysis of multicollinearity for the adequate and correlated data sets used: (a) LARS, (b) Lasso and ElasticNet, (c) Stepwise, (d) Ridge

## 6.4 Analysis of model complexity and stability

The forth stage investigates complexity and stability of the models given by feature selection methods.

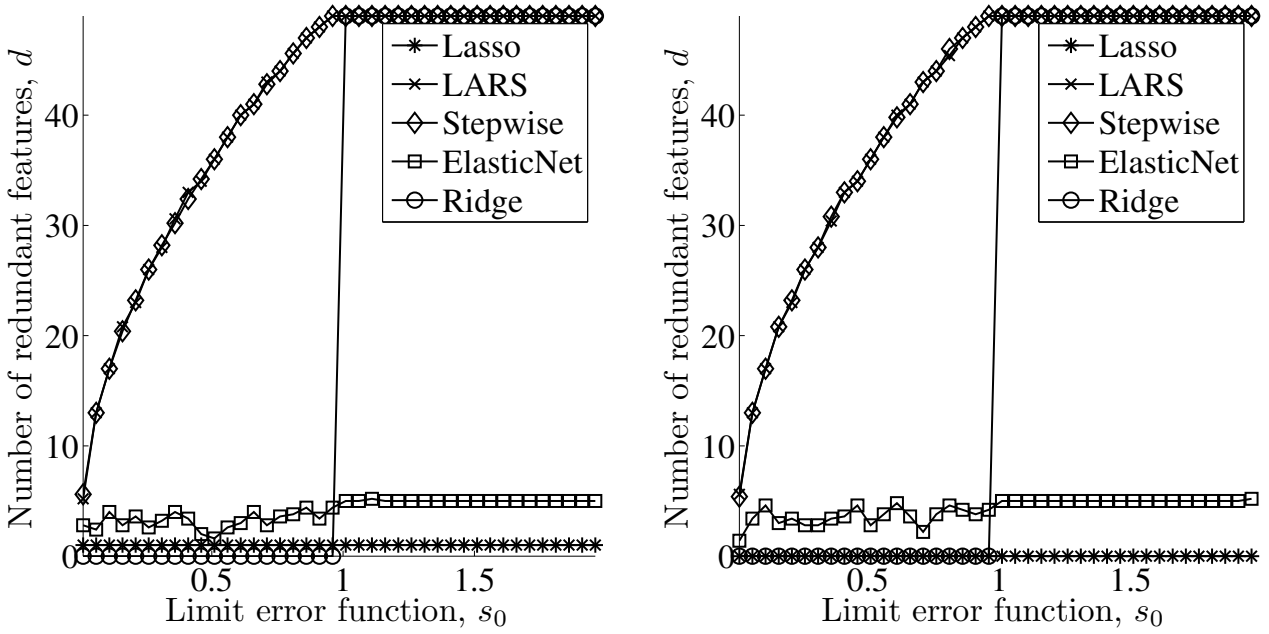
The fig. 11 shows plots of model stability $R = \ln \kappa$ versus the model complexity $C = |\mathcal{A}|$ for the inadequate and correlated data sets since the parameter of multicollinearity $k = 0.2$ and $k = 0.8$. The models are obtained by the illustrative feature selection methods. Every investigated method demonstrates decreasing model stability $R = \ln \kappa$ with the fixed model complexity $C = |\mathcal{A}|$ while the parameter of multicollinearity $k$ rises from 0.2 to 0.8.

The fig. 12 shows the plots of the model stability $R = \ln \kappa$ versus the model complexity $C = |\mathcal{A}|$ for the adequate and redundant data sets since the parameter of multicollinearity $k = 0.2$ and $k = 0.8$. The models are obtained by the illustrative feature selection methods. Lasso gives more stable and less complex model in contrast to other feature selection methods




(a) (b)

Figure 8: Plot of the number of the limit error function $s_0$ versus the redundant features $d$ for the inadequate and correlated data set: (a) $k = 0.2$, (b) $k = 0.8$

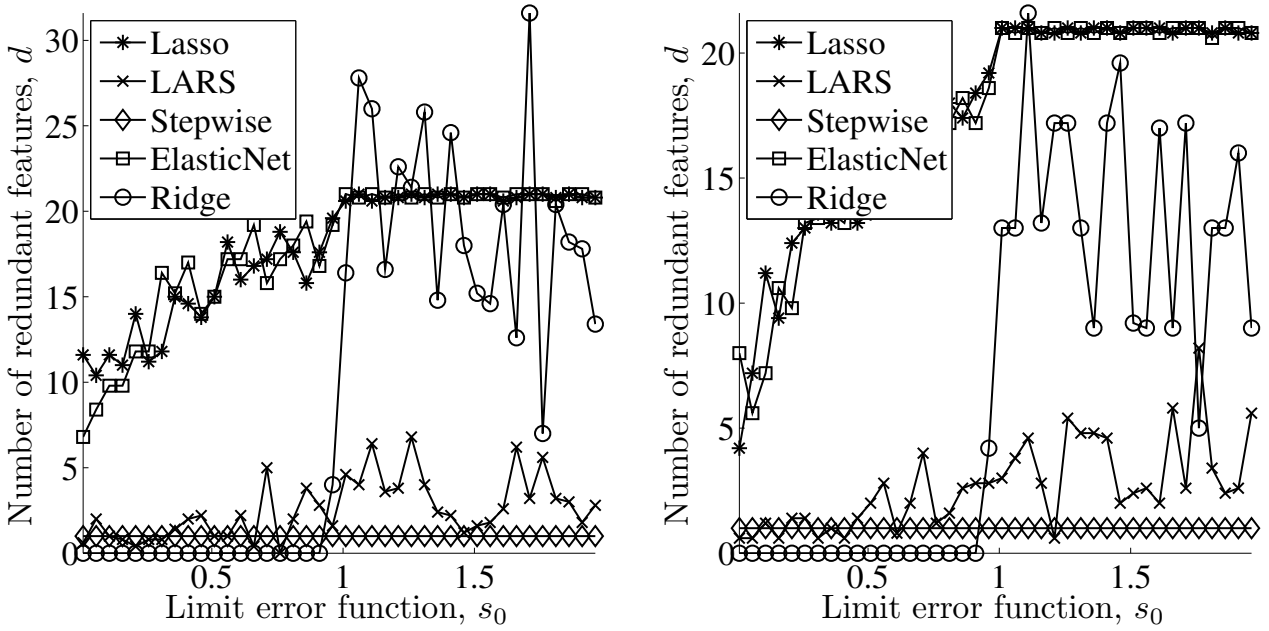



(a) (b)

Figure 9: Plot of the number of the limit error function $s_0$ versus the redundant features $d$ for the adequate and redundant data set: (a) $k = 0.2$, (b) $k = 0.8$

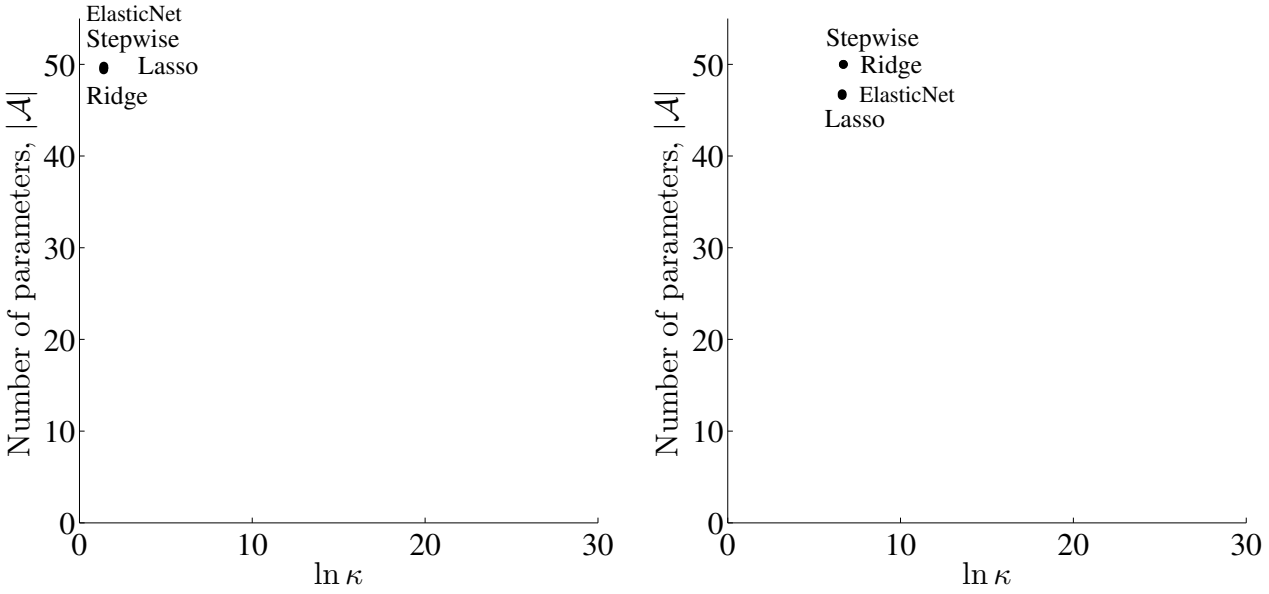
since rising the parameter of the multicollinearity $k$.

The fig. 13 shows the plots of the model stability $R = \ln \kappa$ versus the model complexity $C = |\mathcal{A}|$ for the adequate and correlated data sets since the parameter of multicollinearity $k = 0.2$ and $k = 0.8$. The models are obtained by the illustrative feature selection methods.




(a) (b)

Figure 10: Plot of the number of the limit error function $s_0$ versus the redundant features $d$ for the adequate and correlated data set: (a) $k = 0.2$, (b) $k = 0.8$

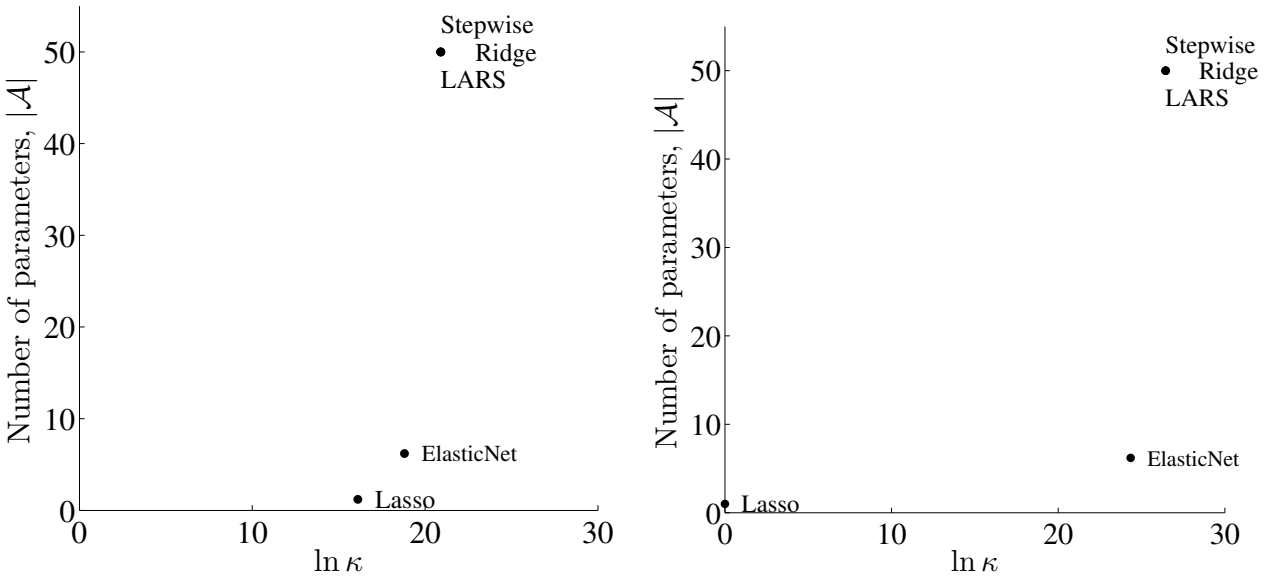



(a) $k = 0.2$ (b) $k = 0.8$

Figure 11: Number of the selected features $|\mathcal{A}|$ and logarithm of the condition number $\kappa$ of the matrix $\mathbf{X}^\intercal\mathbf{X}$ for inadequate and correlated data sets

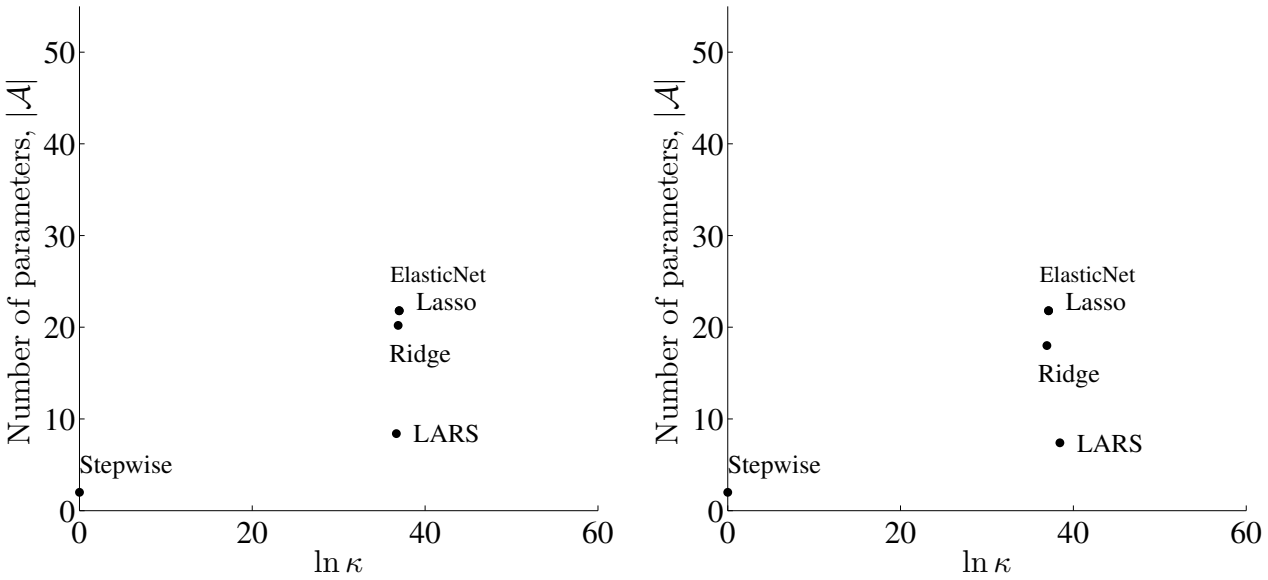
Stepwise gives more stable and less complex model since rising the parameter of multicollinearity $k$ in contrast to other feature selection methods.

The fig. 14 shows the plots of the model stability $R = \ln \kappa$ versus the model complexity $C = |\mathcal{A}|$ for considered data sets. Every point on the fig. 14 corresponds to the some value of the parameter of multicollinearity $k$ from 0.2 to 0.8.




(a) $k = 0.2$ (b) $k = 0.8$

Figure 12: Number of the selected features $|\mathcal{A}|$ and logarithm of the condition number $\kappa$ of the matrix $\mathbf{X}^\intercal\mathbf{X}$ for adequate and redundant data sets

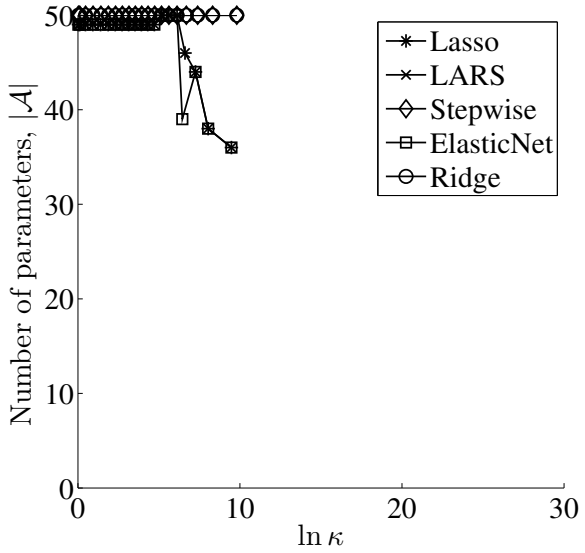



(a) $k = 0.2$ (b) $k = 0.8$

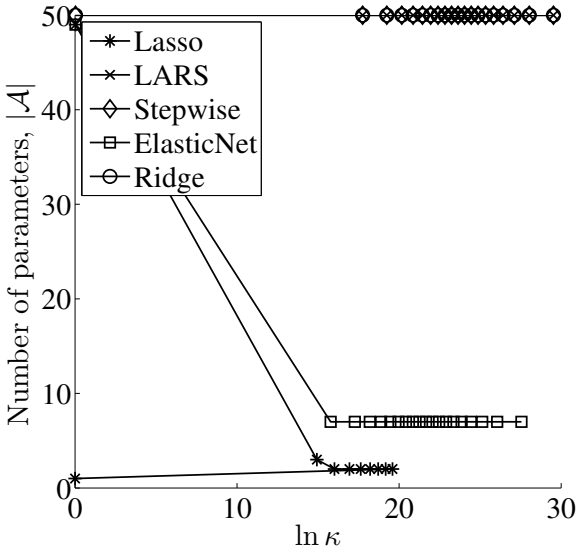
Figure 13: Number of the selected features and logarithm of the condition number $\kappa$ of the matrix $\mathbf{X}^\intercal\mathbf{X}$ for adequate and correlated data sets

# 7 Conclusion

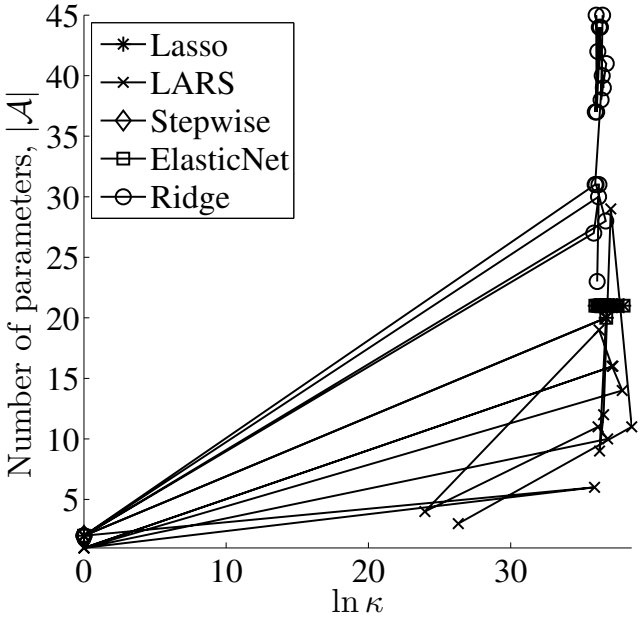
This paper studies the performance of the feature selection methods for data sets of multicollinear features. We propose the test data sets generation procedure to test feature selection methods performance and the criterion of the selected features redundancy to rank feature selection methods by their solution of the multicollinearity problem. Experiments show that Lasso solves the multicollinearity problem for the adequate redundant data sets, Stepwise —




(a) Inadequate correlated data set




(b) Adequate redundant data set




(c) Adequate correlated data set

Figure 14: Complexity and stability of the obtained models using considered feature selection methods since the parameter of multicollinearity $k$ increases from 0.2 to 0.8

for the adequate correlated data sets. None of the considered feature selection methods solves the multicollinearity problem for the inadequate correlated data sets. LARS shows the absence of the relevant features to fit the target vector for the inadequate correlated data sets. The criterion of the selected features redundancy shows that the stable models are given by the same feature selection methods since the parameter of multicollinearity $k$ is small or large. At the same time, the plot of the limit error $s_0$ versus the number of the redundant features $d$ is approximately the same in the one kind data sets since the parameter of multicollinearity $k$ is small or big. All considered methods give unstable models for inadequate correlated data sets, Lasso gives the most stable model for adequate redundant data sets, LARS and Stepwise give

the most stable models for adequate correlated data sets.

The source code of the proposed test data sets generation procedure can be downloaded from [19].

# References


[1] R. G. Askin. Multicollinearity in regression: Review and examples. *Journal of Forecasting*, 1(3):281—292, 1982.

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

[3] David A Belsley, Edwin Kuh, and Roy E Welsch. *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons, New York, 2005.

[4] Lei Yu and Huan Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *ICML*, volume 3, pages 856–863, Washington D.C., 2003.

[5] Rudakov K.V. Strijov V.V., Kuznetsov M.P. Rank-scaled metric clustering of amino-acid sequences. *The Mathematical Biology and Bioinformatics*, 7(1):345–359, 2012.

[6] Yi-Wei Chen and Chih-Jen Lin. Combining svms with various feature selection strategies. In *Feature Extraction*, pages 315–324. Springer, 2006.

[7] George H John, Ron Kohavi, Karl Pfleger, et al. Irrelevant features and the subset selection problem. In *ICML*, volume 94, pages 121–129, 1994.

[8] Vorontsov K. Combinatorial probability and the tightness of generalization bounds. *Pattern Recognition and Image Analysis*, 18(2):243—259, 2008.

[9] Il-Gyo Chong and Chi-Hyuck Jun. Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems*, 78(1–2):103 – 112, 2005.

[10] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157—1182, 2003.

[11] Verónica Bolón-Canedo, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos. A review of feature selection methods on synthetic data. *Knowl. Inf. Syst.*, 34(3):483–519, 2013.

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

[13] M. El-Dereny and N. I. Rashwan. Solving multicollinearity problem using ridge regression models. *Int. Journal of Contemp. Math. Sciences*, 6:585 — 600, 2011.

[14] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[15] B. Efron, T. Hastie, and R. Tibshirani. Least angle regression. *Ann. Statist*, pages 407–499, 2004.

[16] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

[17] Ranjit Kumar Paul. Multicollinearity: Causes, effects and remedies. Technical report, Working paper, unknown date. Accessed Apr. 23, 2013, http://pb8.ru/7hy, 2006.

[18] Strijov Vadim, Krymova Ekaterina, and Weber Gerhard-Wilhelm. Evidence optimization for consequently generated models. *Mathematical and Computer Modelling*, 57(1-2):50–56, 2013.

[19] A. Katrutsa. Source code of the test data sets generation procedure. `http://bit.ly/1qLMyiO`, 2014.