4 - Technological Superiority

Jens Leth Hougaard, *Mette Asmild*

We develop a theoretical framework for analyzing technological possibilities. We consider fundamental properties of technology indexes and demonstrate that previous approaches violate a central axiom dubbed monotonicity in possibilities. From the axiomatic analysis emerge two canonical types of indexes: one based on the volume, and one based on the cardinality of the dominance set. We define a binary superiority relation where both types of indexes have to point in the same direction before concluding that one subset is superior to another.

■ TE-15

Tuesday, 16:00-17:30 - Room 125

Revenue Management with Advertising Applications

Stream: Revenue Management II
Invited session
Chair: *John Turner*

1 - Optimizing Online Advertising Budget Allocation across Multiple Placements

Jian Yang, *Pengyuan Wang*

Big online advertisers are typically faced with a challenging problem in campaign management: how to allocate advertising budget across multiple placements in order to maximize Return on Investment (ROI). We develop a Multi-Touch Attribution (MTA) methodology based on both observation and experimentation to measure ad effectiveness across multiple placements. The MTA empowers a simulator that provides advertisers with what-if analysis for budget allocation. We also build an optimization model using the MTA results to maximize the total ad effectiveness for advertisers, and hence their ROI.

2 - A Class of Nonlinear Allocation Problems with Heterogeneous Substitution

Huaxia Rui, *De Liu*, *Andrew Whinston*

We study the problem of efficiently allocating multiple types of goods (workloads) to multiple agents when different types of goods (workloads) are substitutable and the rates of substitutation differ across agents. We derive theoretical properties of such problems that enable us to design an extremely fast algorithm called SIMS for solving such problems. We expect the SIMS algorithm to work well for real-time applications with time-constrained allocation problems such as the allocation of online advertisement.

3 - The Least Cost Influence Problem

Rui Zhang, *Dilek Gunnec*, *S. Raghavan*

We analyze the diffusion process of a product over a social network while incentives are provided to the individuals. Such catalysation addresses the trade-off of minimizing the amount of incentives given and reaching a greater number of buyers. This problem is NP-Hard for general networks. However, we show that it is polynomially-solvable on tree networks under the assumption that all neighbors of a node exert equal influence. Next, we propose a totally unimodular integer programming formulation based on the insight that the influence propagation network must be a directed acyclic graph.

4 - Foundations of Social Network Ad Optimization

John Turner

We introduce revenue optimization models for placing ads in social networks, motivated by the connectivity structure of the underlying graph. We discuss some pros and cons of the underlying models, and illustrate our approach using real social graphs.

■ TE-16

Tuesday, 16:00-17:30 - Room 127

Model Selection Methods

Stream: Intelligent Optimization in Machine Learning and Data Analysis
Invited session
Chair: *Ivan Reyer*

1 - Multimodelling and Object Selection for Banking Credit Scoring

Alexander Aduenko, *Vadim Strijov*

To construct a bank credit scoring model one must select a set of informative objects (client records) to get the unbiased estimation of the model parameters. This set must have no outliers. The authors propose an object selection algorithm for mixture of regression models. It is based on analysis of the covariance matrix for the parameters estimations. The computational experiment shows statistical significance of the classification quality improvement. The algorithm is illustrated with the cash loans and heart disease data sets.

2 - Comparison of Different Clustering Algorithms Based PCF Classifiers

Emre Çimen, *Gurkan Ozturk*

In this study we dealt with generating different clustering algorithms based polyhedral conic classifiers. The main purpose of using clustering algorithms to generate PCF based classifiers is to determine the number of PCF's and divide the sets to the smaller parts. By this way stronger classifiers can be constructed. Expectation Maximization (EM) and k-Means based algorithms are implemented and tested on well-known literature test problems.

3 - Multicollinearity: Performance Analysis of Feature Selection Algorithms

Alexandr Katrutsa, *Vadim Strijov*

We investigate the multicollinearity problem and its influence on the performance of feature selection methods. The paper proposes the testing procedure for feature selection methods. We discuss the criteria for comparing feature selection methods according to their performance when the multicollinearity is present. Feature selection methods are compared according to the other evaluation measures. We propose the method of generating test data sets with different kinds of multicollinearity. Authors conclude about the performance of feature selection methods if the multicollinearity is present.

4 - Data Mining Application with Decision Tree Algorithms for the Evaluation of Personal Loan Customers' Repayment Performances

Aslı Çalış, *Ahmet Boyacı*, *Kasım Baynal*

Data mining techniques are used extensively in banking area such as many areas. In this study, conducted in banking sector, it was aimed to analysis of available personal loan customers and estimate potential customers' repayment performances with decision tree is one of the classification methods in data mining. In the study, SPSS Clementine was used as a software of data mining. An application was done with C5.0 and C&RT algorithms for evaluation of personal loan customers and the results were compared.

■ TE-17

Tuesday, 16:00-17:30 - Room 005

Conic Optimization and Applications

Stream: Interior Point Methods and Conic Optimization
Invited session
Chair: *Tamás Terlaky*

Multicollinearity: performance analysis of feature selection algorithms

A. Katrutsa, V. Strijov

Moscow Institute of Physics and Technology
Department of control and applied mathematics

IFORS, Barcelona
2014

Research goal

The goal of the research is to develop the procedure to test feature selection methods and propose the criterion to compare feature selection methods through the diagnostic multicollinearity features among selected features.
The problem is that the selected features set contain multicollinearity features and the corresponding model is not stable and simple.
The challenge is to propose a test feature selection method procedure that:

- ranks feature selection methods;
- determines number of multicollinearity features among selected features.

Feature selection problem statement

There given a data set $\mathfrak{D} = \{\mathbf{X}, \mathbf{y}\}$, $\mathbf{X} = [\chi_1, \dots, \chi_n]$ is a design matrix, $j \in \mathcal{J}$, $\mathbf{y} \in \mathbb{R}^m$ is a target vector.
Consider the linear model

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

where the parameters $\mathbf{w} \in \mathbb{W}$, $\mathbb{W}$ is the parameter space and ε is the error vector. The optimum feature subset selection problem is

$$\mathcal{A}^* = \operatorname*{arg\,min}_{\mathcal{A} \subset \mathcal{J}} S(\mathcal{A} | \mathbf{w}^*, \mathfrak{D}_{\mathcal{C}}),$$

where $\mathbf{w}^* = \operatorname*{arg\,min}_{\mathbf{w} \in \mathbb{W}} S(\mathbf{w} | \mathfrak{D}_{\mathcal{L}}, \mathcal{A})$ and $S = \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$.

Consider the following set of the feature selection methods:

$\mathfrak{M} = \{\text{Lasso}, \text{LARS}, \text{Stepwise}, \text{ElasticNet}, \text{Ridge}\}.$

Definitions for multicollinearity and correlation

Definition

The optimum feature subset $\mathcal{A}_i$ *is such* $\mathcal{A}_i \subseteq \mathcal{J}$ *that* $\mathfrak{m}_i : \mathcal{J} \to \mathcal{A}_i$*, where* $\mathfrak{m}_i \in \mathfrak{M}$*.*

Definition

A set of features $\chi_j, \ j \in \mathcal{B}$ *is called multicollinear if there exists* $\delta > 0$ *and coefficients* $a_k, \ k \in \mathcal{B}$ *such that:*

$$\left\| \chi_j - \sum_{k \in \mathcal{B}} a_k \chi_k \right\|_2^2 < \delta,$$

where j *is a feature index and* $j \notin \mathcal{B}$*.*

Definition

A pair of features with index i *and* j *is called correlated if there exists* $\delta > 0$ *such that:*

$$\|\chi_i - \chi_j\|_2^2 < \delta_{ij}.$$

The data sets to test feature selection methods

Inadequate correlated data sets

$$\langle \mathbf{y}, \boldsymbol{\chi}_j \rangle = 0, \quad j \in \mathcal{J};$$

$$\left\| \boldsymbol{\chi}_i - \sum_{l \in \mathcal{B}} \alpha_l \boldsymbol{\chi}_l \right\|_2^2 < \delta,$$

where $\quad i \in \mathcal{J}, \quad i \notin \mathcal{B} \subset \mathcal{J};$

$$\mathcal{J} = \mathcal{P}_y \cap \mathcal{C}_f.$$

Adequate random data sets

$$\mathcal{J} = \mathcal{R}, \ |\mathcal{R}| = r;$$

$$\|\mathbf{y} - \boldsymbol{\chi}_i\|_2^2 < \delta;$$

$$\boldsymbol{\chi}_1, \dots, \boldsymbol{\chi}_r \sim \mathcal{U}[0, 1]^r.$$

Adequate correlated data sets

$$\langle \boldsymbol{\chi}_i, \boldsymbol{\chi}_j \rangle = 0, \quad i, j \in \mathcal{P}_f;$$

$$\|\boldsymbol{\chi}_i - \boldsymbol{\chi}_j\|_2^2 < \delta_{ij}, i \in \mathcal{P}_f, \ j \in \mathcal{C}_f;$$

$$\mathbf{y} = \sum_{j \in \mathcal{P}_f} a_j \boldsymbol{\chi}_j;$$

$$\mathcal{J} = \mathcal{P}_f \cup \mathcal{C}_f.$$

Adequate redundant data sets

$$\left\| \boldsymbol{\chi}_i - \boldsymbol{\chi}_j \right\|_2^2 < \delta_{ij}, \ i, j \in \mathcal{J};$$

$$\|\mathbf{y} - \boldsymbol{\chi}_j\|_2^2 < \delta, \quad j \in \mathcal{J};$$

$$\mathcal{J} = \mathcal{C}_y.$$

Inadequate correlated data sets

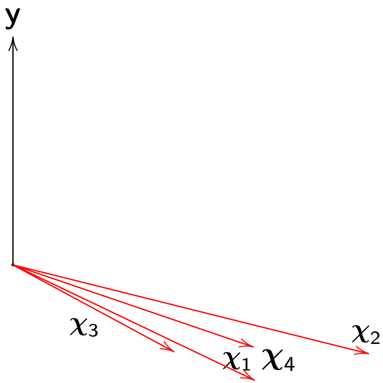
The features with indices from the set $\mathcal{C}_f$ are correlated to each other and the features with indices from the set $\mathcal{P}_y$ are orthogonal to the target vector $\mathbf{y}$.

$$\langle \mathbf{y}, \boldsymbol{\chi}_j \rangle = 0, \quad j \in \mathcal{J};$$

$$\left\| \boldsymbol{\chi}_i - \sum_{s \in \mathcal{B}} \alpha_s \boldsymbol{\chi}_s \right\|_2^2 < \delta,$$

where $\quad i \in \mathcal{J}, \quad i \notin \mathcal{B} \subset \mathcal{J};$

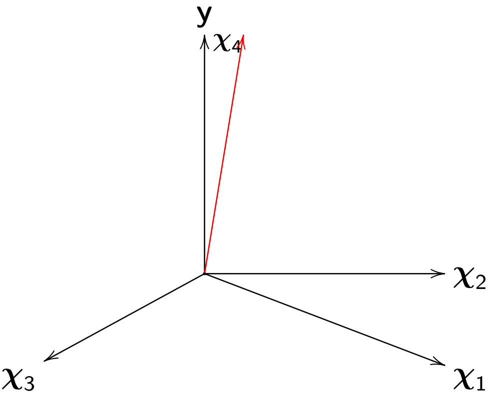
$\mathcal{J} = \mathcal{P}_y \cap \mathcal{C}_f.$



Adequate random data sets

The features with indices from the set $\mathcal{R}$ are generated from the standard uniform distribution and one of the features is correlated with the target vector $\mathbf{y}$.

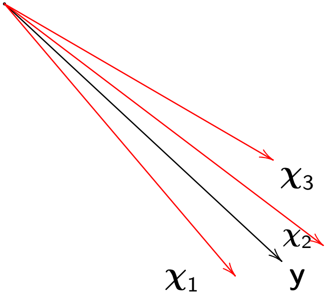
$$\mathcal{J} = \mathcal{R},\ |\mathcal{R}| = r;$$
$$\boldsymbol{\chi}_1, \dots, \boldsymbol{\chi}_r \sim \mathcal{U}[0,1]^r;$$
$$\|\mathbf{y} - \boldsymbol{\chi}_i\|_2^2 < \delta.$$



Adequate redundant data sets

All features with the indices from the set $\mathcal{C}_y$ are correlated with the target vector $\mathbf{y}$.

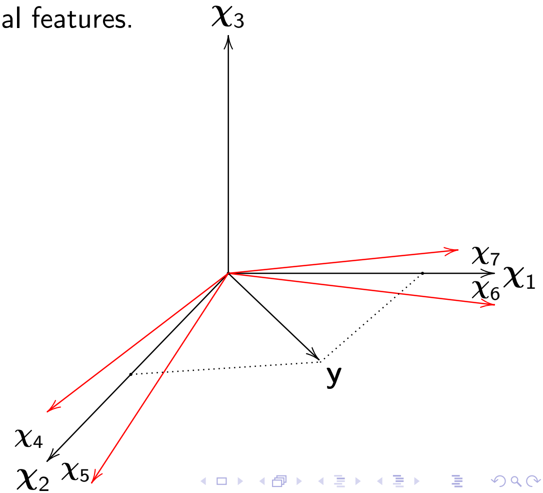
$$\left\|\boldsymbol{\chi}_i - \boldsymbol{\chi}_j\right\|_2^2 < \delta_{ij},\ i,j \in \mathcal{J};$$
$$\|\mathbf{y} - \boldsymbol{\chi}_j\|_2^2 < \delta, \quad j \in \mathcal{J};$$
$$\mathcal{J} = \mathcal{C}_y.$$



Adequate correlated data sets

The set of the orthogonal features with indices from the set $\mathcal{P}_f$, the target vector $\mathbf{y}$ equals some linear combination of the orthogonal features, the set of the features with indices from the set $\mathcal{C}_f$ correlated with the orthogonal features.

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



Data set for generalised test procedure

A test data set structure is specified with following sets:

1) a set of orthogonal features χ_j, where indices $j \in \mathcal{P}_f$;
2) a set of features $\boldsymbol{\chi}_j$ which are orthogonal to the target vector $\mathbf{y}$, where indices $j \in \mathcal{P}_y$;
3) a set of multicollinear features $\boldsymbol{\chi}_j$, where indices $j \in \mathcal{C}_f$;
4) a set of features $\boldsymbol{\chi}_j$ which are correlated to the target vector $\mathbf{y}$, indices $j \in \mathcal{C}_y$;
5) a set of random generated features $\boldsymbol{\chi}_j$, where indices $j \in \mathcal{R}$.

Let k be the multicollinearity parameter: if k equals 1, features are correlated; if k equals 0, features are orthogonal.

The criterion to compare feature selection methods

Let s_0 be some given limit value of error function $S(\mathcal{J}|\mathbf{w}, \mathfrak{D})$. Denote by h the maximum cardinality of the set $\mathcal{J}_h \subseteq \mathcal{A}$ such that the value of the error function is less or equal s_0
$S(\mathcal{J}_h|\mathbf{w}_h, \mathfrak{D}) \leq s_0$

$$h = \mathop{\arg\max}_{S(\mathcal{J}_h|\mathbf{w}_h, \mathfrak{D}) \leq s_0} |\mathcal{J}_h|.$$

In the other words d is the maximum cardinality of the redundant feature set:

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

Criterion

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:

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

Computational experiment

The goal:

- show with the
  1) inadequate correlated
  2) adequate redundant
  3) adequate correlated

  data sets that there is no universal feature selection method according to the proposed criterion;
- show the dependence between the maximum number of the redundant features d and the limit error function value s_0;
- show the dependence between the VIF (Variance Inflation Factor) and the multicollinearity parameter k.

Parameters of the experiment: number of objects $m = 1000$, number of features $n = 50$, the limit error $s_0 = 0.5$, $k = 0.2$ or $k = 0.8$.

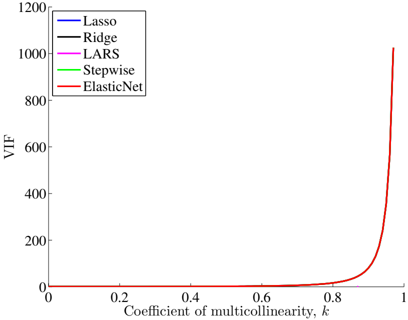
Dependence VIF on the multicollinearity parameter k for inadequate correlated data sets

None of the considered feature selection methods solves the multicollinearity problem for this kind of data sets.

Definition

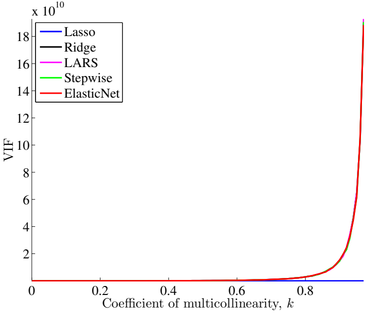
$\mathrm{VIF}_j = \frac{1}{1-R_j^2}$, *where* R_j^2 *is a coefficient of determination, where target vector is* j*-th feature,* $j \in \mathcal{A}, \mathcal{J} = \mathcal{A} \setminus \{j\}$.

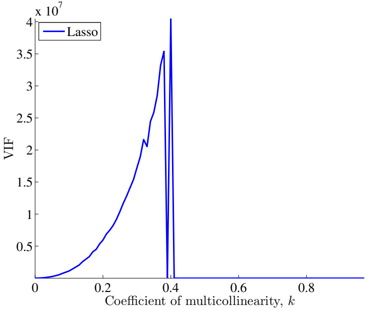
$\mathrm{VIF} = \max\limits_{j \in \mathcal{A}} \mathrm{VIF}_j$



Dependence VIF on the multicollinearity parameter k for adequate redundant data sets

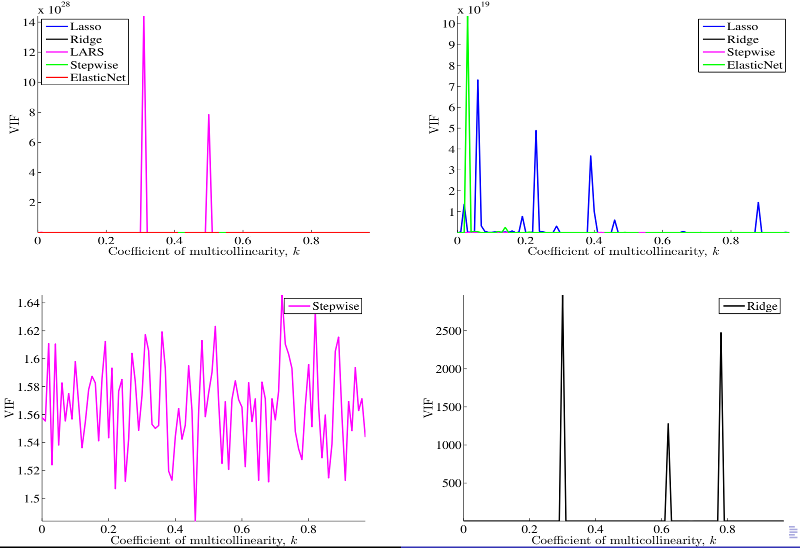
Lasso solves the multicollinearity problem for such kind of data sets.





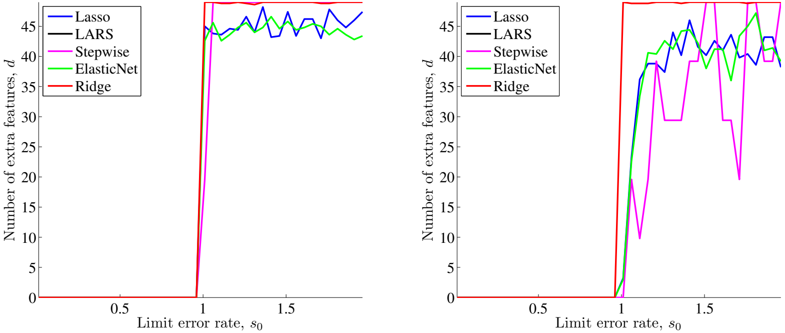
Dependence VIF on *k* for adequate correlated data sets

Stepwise solves the multicollinearity problem.



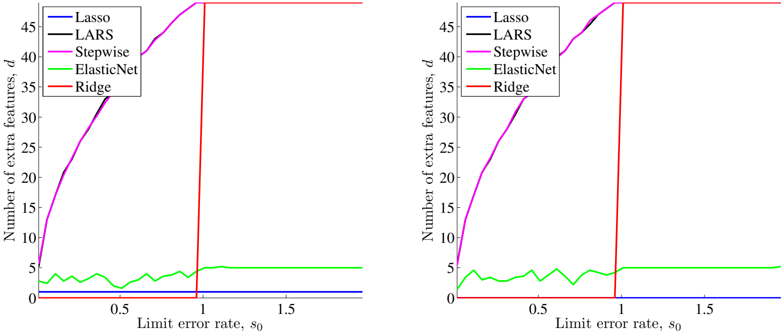
Dependence number of the redundant features d on the limit error s_0 for inadequate correlated data sets

None of the considered feature selection methods gives enough accurate solution to remove features and to stay error function less or equal $s_0 = [0, 1]$.



Dependence number of the redundant features d on the limit error s_0 for adequate redundant data sets

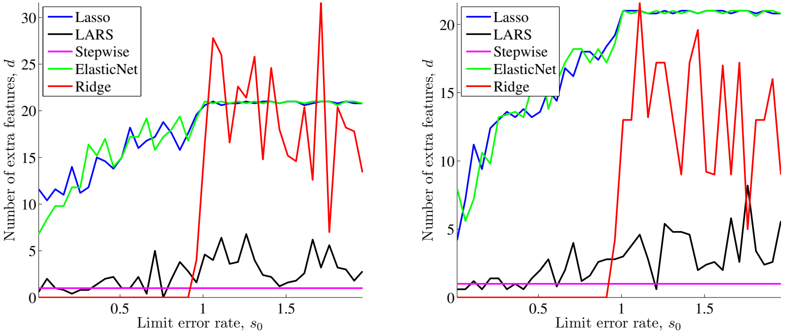
Lasso gives the redundant feature set with the least cardinality.



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

Dependence number of the redundant features d on the limit error s_0 for adequate correlated data sets

Stepwise gives the redundant feature set with the least cardinality.

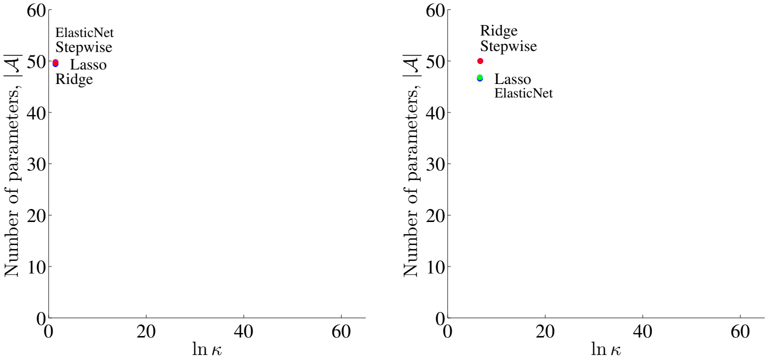


(a) $k = 0.2$

(b) $k = 0.8$

Complexity and stability of the models applied to inadequate correlated data sets

None of the considered feature selection methods gives the stable and simple model.

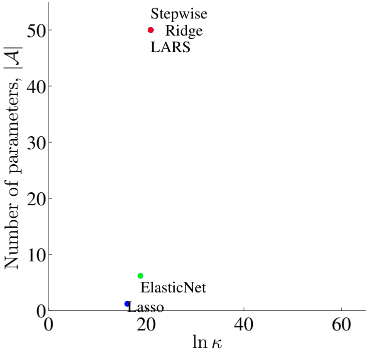


(a) $k = 0.2$

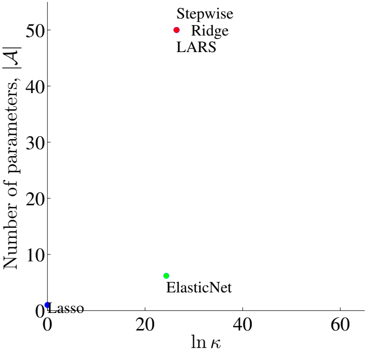
(b) $k = 0.8$

Complexity and stability of the models applied to adequate redundant data sets

With increasing the multicollinearity parameter k Lasso gives more stable and simple model than other methods.



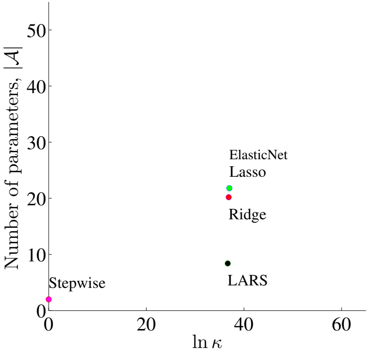
(a) $k = 0.2$



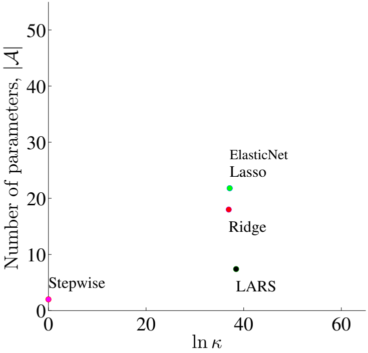
(b) $k = 0.8$

Complexity and stability of the models applied to adequate correlated data sets

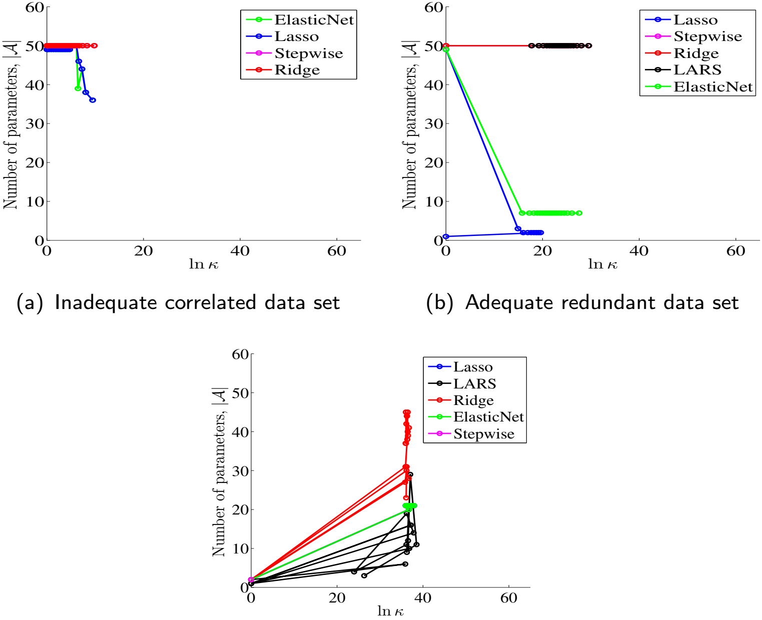
Stepwise gives the most stable and simple model applied to adequate correlated data sets.



(a) $k = 0.2$



(b) $k = 0.8$



(a) Inadequate correlated data set

(b) Adequate redundant data set

(c) Adequate correlated data set

The quality measures of the considered methods

Table: The redundant correlated data set, $k = 0.8$

	d	C_p	RSS	κ	VIF
Lasso	0	$5.16 \cdot 10^8$	$8.5 \cdot 10^{-4}$	1	0.24
Ridge	0	$5.9 \cdot 10^{11}$	0.97	$6.07 \cdot 10^{11}$	$2.9 \cdot 10^9$
Elastic Net	3	$5.16 \cdot 10^8$	$8.5 \cdot 10^{-4}$	$7.3 \cdot 10^{10}$	$2.5 \cdot 10^9$
Stepwise	36	-997	$1.73 \cdot 10^{-12}$	$6.07 \cdot 10^{11}$	$2.9 \cdot 10^9$
LARS	36	-997	$1.65 \cdot 10^{-12}$	$6.07 \cdot 10^{11}$	$2.9 \cdot 10^9$

Table: The adequate correlated data set, $k = 0.8$

	d	C_p	RSS	κ	VIF
Stepwise	1	$9.4 \cdot 10^5$	$8.8 \cdot 10^{-25}$	1	0.63
Ridge	0	$1.8 \cdot 10^{30}$	0.95	10^{16}	$8.65 \cdot 10^{16}$
LARS	1	10^{30}	0.38	$3 \cdot 10^{29}$	10^{20}
Lasso	15	$1.73 \cdot 10^{27}$	$9.2 \cdot 10^{-4}$	$9.92 \cdot 10^{15}$	10^{17}
Elastic Net	15	$1.7 \cdot 10^{27}$	$9.2 \cdot 10^{-4}$	$9.92 \cdot 10^{15}$	10^{17}

Conclusion

- Using various structured data sets we show that there is no universal feature selection method even in the case of linear model.
- We propose the test sets which allow to select the most appropriate feature selection method for any practical application with known structure of data set.
- We develop the test generation procedure to test feature selection methods and select the optimum one by some given criterion.