My first scientific paper

Week 8

Write a peer-review

Vadim Strijov

Moscow Institute of Physics and Technology

2021

Рецензия на статью
Mr. X
«Qwerty»

В работе исследуется метод наименьших квадратов при построении линейных регрессионных моделей. Предлагается вместо метода наименьших квадратов использовать метод наименьших модулей в связи с тем, что сумма модулей разности измерений и соответствующих им значений линейной функции не является всюду дифференцируемой. Приводятся формулы расчета коэффициентов одномерной линейной модели по МНК. Для получения робастных коэффициентов предлагается использовать взвешенный МНК с весами вида $1/|y|^q$. Указан критерий оптимальности значений q. Адекватность полученной модели проверяется с помощью критерия Фишера. В качестве примера использования предложенного метода приведена выборка из пяти элементов. По результатам работы сделаны выводы.

Рецензируемая статья не содержит ни аннотации, ни введения. В статье не сообщаются целей работы. Статья внутренне противоречива: несколько раз предлагается использовать метод наименьших модулей (стр.1, абз. 1.; стр. 3, абз. 6, 7; стр. 4., абз. 1, 2.), однако для отыскания коэффициентов линейной модели использует метод наименьших квадратов. Следует отметить, что для нахождения весов линейной модели при минимизации суммы модулей разностей не имеет смысл дифференцировать (2). Для этих целей используются, например, методы линейного программирования.

Тематика отыскания робастных линейных моделей с использованием функционала (2), поднимаемая в рецензируемой статье, подробно освящена, например, в работе ...

В связи с вышеизложенным, считаю, что рецензируемую статью «Qwerty» публиковать в журнале «Journal» необязательно.

Рецензент,
к.ф.-м.н. Mr. Y

Рецензия на статью Mr. X

«QWERTY»

В статье описана линейная регрессия одной переменной – высоты на плотность почвы. Рассмотрены три выборки; одна состоит из шести элементов, две другие содержат по три элемента. Приведены коэффициенты трех линейных функций регрессии.

В программу конференции … входит рассмотрение фундаментальных математических вопросов распознавания, интеллектуального анализа данных, машинного обучения, прогнозирования, прикладных задач и программных систем. Математическая часть рецензируемой статьи опирается на книгу Ю.В. Линника, переизданную в 1962 году. Фундаментальная часть метода изложена на стр. 11 этой книги и проиллюстрирована похожим примером из работы Д.И. Менделеева 1881 года. На стр. 20 книги приведен обзор теоретических и прикладных работ известных исследователей за 1806—1946 годы, посвященных решению рассматриваемой в рецензируемой статье задачи.

В связи с вышеизложенным рецензент не считает уместным поднимать данную задачу для повторного обсуждения ее математического аппарата на конференции … и предлагает авторам подать статью на конференцию, посвященную вопросам почвоведения.

Рецензент,

к.ф.-м.н., доц. Mr. Y

Рецензия на статью Mr. X
«Qwerty»

В статье рассмотрена весьма актуальная проблема построения линейных структурных соотношений между случайными величинами на малых выборках. На практике проблема восстановления закономерностей на малых выборка часто связана с высокой стоимостью экспериментов и может встать очень остро. В современной литературе предлагаются, по крайней мере, три основных подхода: 1) ведение специальных функций ошибки (или функций качества) модели, 2) отказ от сильных гипотез порождения данных (использование достаточно общей информации о законах распределения исследуемых случайных величин) и 3) восстановление совместного распределения входных и зависимых случайных величин. Авторы выбрали второй путь и рассмотрели практически важные случаи одномерного и многомерного линейного структурного соотношения, а также доказали теорему о несмещённости получаемых оценок параметров. Также в статье был поставлен вычислительный эксперимент на синтетических данных: выборки различного объема были порождены согласно экспоненциальному, логнормальному, усеченному нормальному распределению и распределению Рэлея; получены хорошие результаты, которые сравнивались с ранее предложенными.

Статья полезна, аккуратно написана, содержит интересный результат и хороший вычислительный эксперимент. Предлагаю опубликовать статью в <Journal> без доработок.

Рецензент
к.ф.-м.н. Mr. X

14 июля 2012 г.

Name of paper

1. The introduction should carry the brief explanation what is the Operating Theater Layout and the activity. If is difficult to read the massage without knowing the main subjects.

2. The introductory parts (1..3) are too long. If in one-page text the goal, the novelty and the importance will be explained, it would be good.

3. Problem statement and problem modeling should be joined; the problem statement should be reduced to the main message.

4. Parts 5, 6. Please write, what doctors (users) say about this placement: what kind of placement is better: algorithmic or manual and according to what criterion?

Methodologies and Tools to ...

1. The abstract must convey the field and the main problem of the investigation. Now the abstract is a part of the introduction.

2. It would be great to eliminate the vague sentences like "The increasing globalization of markets" from the introduction and write about the goals and the novelty of the paper. The main subject of the paper, NPD, must go first.

3. Part 2. It would be great if the text and the table will be tightly connected. The table is the key here.

4. Part 3 is the main part of the paper; is too brief. It should answer to the following questions:

What is the source of the document collection?

What are the selection criterions?

Why the authors consider the criterions to be adequate to the goal of the investigation?

How the percentage was calculated?

How the graphs were constructed?

What conclusion the reader could make from the figures?

Item 1..9: could the percent be shown as a histogram?

5. The conclusion repeats the previous part. If it will deliver how the reader can use the results in his practice, it would be good.

Comprehensive study of feature selection methods ~~to~~for solving~~e~~ the multicollinearity problem according to evaluation criteria[1][2][3][4]

This[5][6] paper provides a new approach ~~for the~~to feature selection~~. It is~~ based on the concept of feature filters, so ~~the~~that feature selection is independent of the prediction model. Data fitting is stated as a single-objective optimization problem, where the objective function indicates the error of approximating~~ion~~ the target vector ~~with~~as some function of given features. ~~The l~~Linear dependence between features ~~indicates~~induces the multicollinearity problem~~. It~~ and leads to ~~un~~instability of the model and redundancy of the feature set. Thise paper introduces a feature selection method based on ~~a~~ quadratic programming ~~approach~~. This approach takes into account the mutual dependence of the features and the target vector, and selects features according to relevance and similarity measures~~, which are~~ defined according to ~~an application~~the specific problem. The main idea is to minimize mutual dependence and maximize approximation quality by varying a binary vector~~,~~ that indicates~~ing~~ the presence of features~~ presence~~. The selected model is less redundant and more stable. To evaluate the quality of the proposed feature selection method and compare it with others, we use several criteria to measure ~~un~~instability and redundancy. In ~~the~~our experiments, we compare the proposed approach with ~~the~~several other feature selection methods~~: LARS, Lasso, Ridge, Stepwise and Genetic algorithm. We~~, and show that the quadratic programming approach gives superior results according to the criteria considered ~~criteria on~~for the test and real data sets.

1 Introduction

~~–~~This paper presents a ~~novel~~new approach to avoiding multicollinearity in feature selection. *Multicollinearity* is a strong correlation between features~~,~~ ~~which~~that affect the target vector simultaneously. ~~Due to~~In the presence of multicollinearity, ~~the~~ common methods of regression analysis, ~~like~~such as least squares, build unstable models of excessive complexity. The formal definitions of model stability, complexity and redundancy are given in Section 5.

Most ~~of previously proposed~~existing feature selection methods that solve the multicollinearity problem are based on ~~various~~ heuristics [Leardi (2001), Oluleye et~~-~~ al. (2014)~~Oluleye, Armstrong, Leng Diepeveen~~], greedy searches [Ladha and Deepa (2011), Guyon (2003)] or regularization techniques [Zou and Hastie (2005), El-Dereny and Rashwan (2011)]. These approaches do not take into account the data set configuration and do not guarantee the optimality of the specially designed feature subset [Katrutsa and Strijov (2015)]. In contrast, we propose ~~to use~~a *quadratic programming approach* [Rodriguez-Lujan et~~-~~ al. (2010)~~Rodriguez-Lujan, Huerta, Elkan Cruz~~] to solv~~ing~~e the multicollinearity problem that ~~corrects~~ avoids the disadvantages mentioned above. This approach is based on two ideas: ~~the first one is to~~ represent~~ing~~ feature presence as a binary vector, and ~~the second one is to~~ defin~~ing~~e the feature subset quality criterion in ~~a~~ quadratic form. The first term of the quadratic form is the pairwise feature similarit~~y~~ies, and the linear term is the relevance of feature~~s~~ ~~relevances~~ to the target vector. Therefore, we can state the feature selection problem with ~~the~~a quadratic objective function and a ~~b~~Boolean vector domain.~~–~~

Measures of feature similarit~~y~~ies and relevance~~s~~ are problem-dependent and ~~have~~need to be defined ~~before performing feature selection~~ according to the application ~~problem~~before performing feature selection. These measures ~~have to~~should take into account the data set configuration to remove redundant, noisy and multicollinear features, selecting those~~,~~ ~~which~~that are significant for target vector approximation. We consider the correlation coefficient [Hall (1999)] and ~~the~~ mutual information [Estaez et~~-~~ al. (2009)~~Estaez, Tesmer, Perez Zurada~~] between features as measures of feature similarit~~y~~ies ~~as well as~~and between features and the target vector as a measure of feature relevance~~s~~. These measures guarantee ~~the~~a positive semidefinite quadratic form.~~–~~

To solve the *convex optimization problem* we need to relax the binary domain to ~~the~~a continuous ~~one~~domain. ~~After t~~This relaxation, allows ~~we have~~ the *convex optimization problem*~~,~~ ~~which can~~ to be efficiently solved by state-of-the-art solvers~~,~~ ~~for example from~~such as CVX, a package for specifying and solving convex programs ~~package by~~ [Grant and Boyd (2014), Grant and Boyd (2008)]. To ~~return from~~translate the continuous solution to ~~the~~a binary ~~one~~solution, we set a *significance threshold*~~,~~ ~~which~~that defines a number of features to be selected ~~features~~. If the feature similarity function does not give a positive semidefinite matrix, then the optimization problem is not convex~~,~~ and convex relaxation is required. In this case, ~~the authors~~we propose ~~to use the~~using a semidefinite programming relaxation ~~by~~ [Naghibi et~~-~~ al. (2015)~~Naghibi, Hoffmann Pfister~~]. Such feature similarity functions are out of the scope of this paper. In addition, the proposed approach gives a simple visualization of the feature weights in the target vector approximation. This visualization helps to tune the threshold.

We ~~carry out~~perform experiments on special test data sets generated according to the procedure proposed in [Katrutsa and Strijov (2015)]. These data sets demonstrate different cases of multicollinearity between features and correlation between features and the target vector. Experiments show that the proposed approach outperforms the other ~~considered~~ feature selection methods considered on every type of test data set~~s~~. ~~Also, q~~Quadratic programming feature selection ~~shows~~also gives better quality results on the test and real data sets according to various simultaneous evaluation criteria ~~simultaneously~~ in contrast to other feature selection methods.

The main contributions of this paper are:

- ~~It~~ addressing~~es~~ the multicollinearity problem with a quadratic programming approach and investigating~~es~~ its properties~~y~~;~~.~~
- ~~It demonstrates~~evaluating the performance of the quadratic programming feature selection method on ~~the~~ test data sets according to various criteria;~~.~~
- ~~It~~ comparing~~es~~ the proposed feature selection method with other~~s~~ methods on test and real data sets, and showing~~s~~ that ~~it~~the proposed method gives ~~the~~ better feature subsets than the other methods. The feature subset quality ~~are~~is measured by external criteria.

Related ~~works~~research

A comprehensive survey of feature selection algorithms ~~was~~ can be found in [Li et~~-~~ al. (2016)~~Li, Cheng, Wang, Morstatter, Trevino, Tang — Liu~~],~~.~~ ~~It~~which gives a systematic analysis ~~for~~of filter, wrapper, and embedded methods.~~ A number of algorithms are collected in library~~[1]. ~~Previously, v~~Various strategies ~~were~~have been proposed ~~to~~for detecting multicollinearity and ~~to~~ solving~~e~~ ~~this~~the multicollinearity problem [Askin (1982), Leamer (1973), Belsley et~~-~~ al. (2005)~~Belsley, Kuh — Welsch~~]. One way to solve the multicollinearity problem is to use feature selection methods [Liu and Motoda(2012), Belsley et~~-~~ al. (2005)~~Belsley, Kuh — Welsch~~]. These~~y~~ are based on ~~some~~ scoring functions~~,~~ ~~which~~that estimate the quality of a feature subset, or on ~~some~~a heuristic sequential search procedure.

This paper considers feature selection methods~~, which are~~ based on scoring functions, ~~like~~such as least angle regression (LARS) [Efron et~~-~~ al. (2004)~~Efron, Hastie, Johnstone, Tibshirani et al.~~], ~~L~~lasso [Tibshirani (1994)], ~~R~~ridge regression [El-Dereny and Rashwan (2011)], and the ~~E~~elastic ~~N~~net [Zou and Hastie (2005)], and ~~which are~~ based on ~~the~~ sequential search, ~~like~~such as ~~S~~stepwise regression [Harrell (2001)] and the ~~G~~genetic algorithm [Ghamisi and Benediktsson (2015)]. The ~~L~~lasso scoring function is the weighted sum of the ℓ_2 norm of the residuals and the ℓ_1 norm of the parameter vector. This scoring function gives a good approximation ~~of~~to the target vector and penalizes ~~big~~large elements in the parameter vector. Moreover, the ℓ_1 norm of the parameter vector induces sparsity ~~of~~in the obtained parameter vector and therefore performs feature selection. The ~~R~~ridge scoring function is the same as in ~~L~~lasso, but uses the ℓ_2 norm instead of the ℓ_1 norm~~, it uses~~ ~~ℓ_2 norm~~. This approach makes the solution more stable, but does not give a sparse parameter vector and selects features ~~not so~~less aggressively ~~as~~than ~~L~~lasso. The ~~E~~elastic ~~N~~net [Zou and Hastie (2005)] uses a linear combination of the ℓ_1 and ℓ_2 norms of the parameter vector as a penalty ~~to~~for the residual norm. This penalty allows us to combine~~ing~~ the advantages of both ~~L~~lasso and ~~R~~ridge regression~~methods~~. ~~The~~Two common problems for these ~~mentioned~~ feature selection methods are ~~how to~~ tuning~~e~~ the weights corresponding to the penalty terms and ~~how to~~ taking~~e~~ into account the structure of a data set. ~~Another group of~~A study of feature selection methods that use~~performs~~ sequential search can be found in [Aha and Bankert (1996)]. The ~~G~~genetic algorithm [Ghamisi and Benediktsson (2015)] uses a random search that

[1] Implementations of several feature selection algorithms are available from a library developed by Arizona State University (http://featureselection.asu.edu).

maximizes the objective function and adds or removes some ~~number of~~ features on ~~every~~each iteration,~~. On the other hand,~~ while ~~S~~stepwise regression starts from ~~the~~an empty feature set and sequentially adds a single feature on ~~every~~each interation according to the importance ~~obtained by performing~~determined by an F-test.

2 Feature Selection Problem Statement

Let $\mathbf{X}=[\chi_1,\dots,\chi_n]\in\mathbb{R}^{m\times n}$ be ~~the~~a design matrix, where $\chi_j\in\mathbb{R}^m$ is the j-th feature. ~~Let $\mathbf{y}\in\mathbb{R}^m$ be the target vector.~~ Denote by $\mathsf{J}=\{1,\dots,n\}$ the feature index set,~~.~~ and ~~L~~let $\mathsf{A}\subseteq\mathsf{J}$ be a feature index subset. Let $\mathbf{y}\in\mathbb{R}^m$ be a target vector. The data fitting problem is to find a parameter vector $\mathbf{w}^*\in\mathbb{R}^n$ such that

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

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

This study explores the linear function

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

where $\mathbf{X}_{\mathsf{A}}$ is the reduced design matrix~~, which~~ consisting~~s~~ of features with indices ~~from set~~in A, and the quadratic error function

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

The ~~F~~features ~~$\chi_j, j\in\mathsf{J}$~~ are ~~supposed~~assumed to be noisy, irrelevant or multicollinear,~~. It~~ which leads to ~~an~~ additional error in estimating~~ion of~~ the optimum vector $\mathbf{w}^*$ and increases the ~~un~~instability of this vector. ~~One can use f~~Feature selection methods can be used to remove ~~named~~certain features from the design matrix $\mathbf{X}$. The feature selection procedure reduces the dimensionality of problem (1) and improves the stability of the optimum vector $\mathbf{w}^*$. The feature selection problem is

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

~~-~~where $Q:\mathsf{A}\to\mathbb{R}$ is a quality criterion~~, which~~ that ~~validates~~determines the quality of ~~some~~a selected feature index subset $\mathsf{A}\subseteq\mathsf{J}$. Problem (3) does not necessarily require ~~any~~ estimation of the optimum parameter vector $\mathbf{w}^*$. It uses the relationships between the features $\chi_j, j\in\mathsf{J}$ 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 $j\in\mathsf{A}$. ~~So~~Then problem (3) can be rewritten~~:~~ as

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

~~-~~where $Q:\mathbb{B}^n\to\mathbb{R}$ is another form of the criterion Q with domain $\mathbb{B}^n$. The ~~V~~vector $\mathbf{a}^*$ and the index set A^* are ~~corresponding as~~related by

$$a_j^*=1\Leftrightarrow j\in\mathsf{A}^*,\ j\in\mathsf{J}.\qquad(5)$$

2.1 Multicollinearity problem

In this subsection, we give a formal definition and some special cases of the multicollinearity problem~~phenomenon and special cases~~. 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 an active index subset $\mathcal{A} \subseteq \mathcal{J}$.

Definition 2.1 *The features with indices ~~from~~ in the set $\mathcal{A}$ are ~~called~~ multicollinear if there exist an index j, coefficients λ_k, an index $k \in \mathcal{A} \setminus j$ and a 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 the *degree of multicollinearity*.

~~The~~A particular case of this definition is the following.

Definition 2.2 *~~Let t~~The features indexed by i, j ~~be~~are correlated if there exists a 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 $\lambda_k = 0, k \neq j$ and $\lambda_k = 1, k = j$.

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

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

3 Quadratic Optimization Approach to the Multicollinearity Problem

~~The paper~~In [Katrutsa and Strijov (2015)], it was shown~~s~~ that none of the ~~considered~~ feature selection methods considered (LARS, ~~L~~lasso, ~~R~~ridge regression, ~~S~~stepwise regression and the ~~G~~genetic algorithm) solve ~~the~~ problem (1) and give a model that is simultaneously stable, accurate and nonredundant ~~model simultaneously~~. Therefore, we propose ~~the~~a quadratic programming approach to solv~~e~~ing the 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 problem (4) ~~in the form of~~ as a quadratic function

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

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

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

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

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

Definition 3.1 *Let QP(Sim, Rel) be a feature selection method~~, which~~ that solves the optimization problem*

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

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

$$\mathbf{Q} = [q_{ij}] = \mathrm{Sim}(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j) \text{ [7]}, \tag{12}$$

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

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

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

3.1 Correlation coefficient

The similarit~~ies~~y between the features $\boldsymbol{\chi}_i$ and $\boldsymbol{\chi}_j$ can be computed ~~with~~using the Pearson correlation coefficient [Hall (1999)]. The Pearson 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 the covariance between ~~features~~ $\boldsymbol{\chi}_i$ and $\boldsymbol{\chi}_j$, and $\mathrm{Var}(\cdot)$ is the variance of a feature. The sample correlation coefficient is defined as

$$\hat{\rho}_{ij} = \frac{(\boldsymbol{\chi}_i - \bar{\boldsymbol{\chi}}_i)\,(\boldsymbol{\chi}_j - \bar{\boldsymbol{\chi}}_j)}{\|\boldsymbol{\chi}_i - \bar{\boldsymbol{\chi}}_i\|_2 \|\boldsymbol{\chi}_j - \bar{\boldsymbol{\chi}}_j\|_2}, \qquad \bar{\boldsymbol{\chi}}_i = [\bar{\chi}_i, \ldots, \bar{\chi}_i], \qquad \bar{\boldsymbol{\chi}}_j = [\bar{\chi}_j, \ldots, \bar{\chi}_j] \tag{14}$$

where $\bar{\chi}_i$ and $\bar{\chi}_j$ are the means of ~~features~~ $\boldsymbol{\chi}_i$ and $\boldsymbol{\chi}_j$ respectively. In this case, the elements of ~~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 ~~vector~~ $\mathbf{b} = [b_i]$ are equal to the absolute values of the sample correlation coefficient between the feature $\boldsymbol{\chi}_i$ and the target vector $\mathbf{y}$:

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

~~It~~This 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

~~The~~An alternative measure of feature similarity ~~measure~~ is based on the concept of mutual information ~~concept~~ [Estaez et. al. (2009)~~Estaez, Tesmer, Perez Zurada~~, Peng et. al. (2005)~~Peng, Long Ding~~]. The mutual information between the features χ_i and χ_j is defined as

$$I(\chi_i,\chi_j)=\iint p(\chi_i,\chi_j)\log\frac{p(\chi_i,\chi_j)}{p(\chi_i)p(\chi_j)}d\chi_i d\chi_j. \quad (17)$$

The sample mutual information is calculated based on an estimation of the probability distribution in equation (17). To estimate the marginal and joint probability distributions, we use the approach described in Section 4.1. of ~~the paper~~ [Peng et. al. (2005)~~Peng, Long Ding~~]. ~~In t~~This ~~paper, authors~~approach uses the Parzen window method with a Gaussian kernel to estimate the probability distributions, which are necessary for computing the mutual information~~ computation~~, and replaces~~ing~~ integration ~~for~~with summation to compute the mutual information.

In this case, the elements of ~~matrix~~ $\mathbf{Q}=[q_{ij}]$ are equal to the values of the corresponding sample mutual information:

$$q_{ij}=\mathrm{Sim}(\chi_i,\chi_j)=I(\chi_i,\chi_j),$$

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

$$b_i=\mathrm{Rel}(\chi_i)=I(\chi_i,\mathbf{y}).$$

3.3 Normalized feature significance

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

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

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

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

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

~~Thus, to represent the feature relevance w~~We propose ~~to use in (13)~~using the normalized feature significance to represent feature relevance:

$$b_j=\mathrm{Rel}(\chi_j)=\hat{p}_j. \quad (19)$$

3.4 Convex representation of the feature selection problem

The quadratic programming approach to the multicollinearity problem leads to problem

(11), which is NP-hard due to itsbecause of the bBoolean domain. Therefore, we need to approximate itthis problem with thea convex optimization problem to solve it efficiently.

Assume that function Sim gives thea positive semidefinite matrix $\mathbf{Q}$., tThen the quadratic form (9) is thea convex function. To represent problem (11) in the convex form, we have to replace the non-convex set B^n with thea convex oneset. TheA natural way for this representationto achieve this is to use the convex hull of set B^n:

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

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

$$\mathbf{z}^* = \arg\min_{\mathbf{z}\in[0,1]^n} \mathbf{z}^T\mathbf{Q}\mathbf{z} - \mathbf{b}^T\mathbf{z} \quad \text{s.t.} \quad \|\mathbf{z}\|_1 \leq 1. \tag{20}$$

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

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

Definition 3.3 *LetThe value* τ *beis a significance threshold such thatif* $z_j^* > \tau$ *if and only if* $a_j^* = 1$ *and* $j \in \mathsf{A}^*$.

–Tuning the value of τ is problem-dependent and is based on the appropriate error rate, the number of selected features selected and the values of the evaluation criteria. To obtain the most appropriate significance threshold for a specific problem, One haswe need to set somea range of values for τ to get the most appropriate one for considered problem. In Section 6, we showpresent some 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 [Katrutsa and Strijov (2015)]. These data sets to demonstrate the performance of several feature selection methods in the multicollinearity problem. Below wWe provide a summary of these data sets below.

Definition 4.1 *LetAn inadequate and correlated data set be a data set that consists of the correlated features, which that are orthogonal to the target vector, (Fig. 1).*

Definition 4.2 *LetAn adequate and random data set be a data set that consists of the random features with theand a single feature, which that approximates the target vector, (Fig. 2).*

Definition 4.3 *LetAn adequate and redundant data set be a data set that consists of the features, which that are correlated to the target vector, (Fig. 3).*

Definition 4.4 *LetAn adequate and correlated data set be a data set that consists of the orthogonal features and features, that are correlated to the orthogonal onesfeatures;. Tthe target vector is athe sum of two orthogonal features, (Fig. 4).*

The performances of the considereddifferent feature selection methods isare compared according tousing various evaluation criteria, which are provided in the next section.

5 Evaluation Criteria

To evaluate ~~a~~the quality of ~~the~~a selected feature subset and to compare ~~considered~~ feature selection methods, we use the following criteria ~~used in papers~~from [Paul (2006), Paul and Das (2015)].

Variance inflation factor. To ~~diagnose~~detect multicollinearity, ~~the paper~~ [Paul (2006)] uses the variance inflation factor, VIF_j. ~~It~~which shows ~~the~~any linear dependence between the j-th feature and the other features. To compute VIF_j, we estimate the parameter vector $\mathbf{w}^*$ according to problem (1) assuming that $\mathbf{y} = \boldsymbol{\chi}_j$, and extract~~ing~~ the j-th feature from ~~index set~~ $\mathsf{A} = \mathsf{A} \setminus j$:

$$VIF_j = \frac{1}{1 - R_j^2},$$

where $R_j^2 = 1 - \frac{RSS_j}{TSS_j}$ is the coefficient of determination, ~~and~~

$$RSS_j = \left\|\boldsymbol{\chi}_j - \mathbf{X}_{\mathsf{A}}\mathbf{w}^*\right\|_2^2, \qquad TSS_j = \left\|\boldsymbol{\chi}_j - \overline{\boldsymbol{\chi}}_j\right\|_2^2,$$

~~where~~and $\overline{\boldsymbol{\chi}}_j$ is defined in (14). ~~The paper~~In [Paul (2006)], ~~states~~it is stated that if $VIF_j \geq 5$ 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~~over all $j \in \mathsf{A}$:

$$VIF = \max_{j \in \mathsf{A}} VIF_j.$$

Stability. To estimate the stability R of ~~parameters~~ $\mathbf{w}^*$ estimated on a selected feature subset A, we use the logarithm of the ~~inverse~~ reciprocal of the condition number of ~~matrix~~ $\mathbf{X}^T\mathbf{X}$ ~~,~~:

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

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

Complexity. To measure the complexity C of a selected feature subset A^*, we use the cardinality of ~~this subset~~ A^*:

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

~~The less~~A smaller complexity ~~is, the~~value corresponds to better subset selection~~ed subset~~.

Mallow's C_p. ~~The~~ Mallow's C_p criterion [Gilmour (1996)] is a trade~~s-~~-off between the residual norm $r = \left\|\mathbf{y} - \mathbf{X}\mathbf{w}\right\|_2^2$ and the number of features p. ~~The~~ Mallow's C_p is defined as

$$C_p = \frac{r_{\mathsf{A}}}{r} - m + 2p,$$

where $r_{\mathsf{A}} = \left\|\mathbf{y} - \mathbf{X}_{\mathsf{A}}\mathbf{w}\right\|_2^2$ is computed ~~with~~using $p = |\mathsf{A}|$ features ~~only~~ and m is the number of rows in the design matrix, which is the same for matrices and in both $\mathbf{X}$ and $\mathbf{X}_A$. In terms of this criterion, thea smaller value for C_p is, theindicates a better feature subset.

Bayesian information criterionBIC. The Bayesian Iinformation criterion — BIC [McQuarrie and Tsai (1998)] is defined as

$$BIC = r + p \log m.$$

The notation here is the same as in the definition of Mallow's C_p criterion definition. TheA smaller value offor BIC is, theshows a better fit between the model fitsand the target vector. ConsideredThe criteria are summarized in the Table 1.

7 Conclusion

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