Математички Билтен
41(LXVII) No. 1
2017(80-100)
Скопје, Македонија

ISSN 0351-336X (print)
ISSN 1857-9914 (online)
UDC: 519.246:519.71

REGULARIZED LEAST-SQUARE OPTIMIZATION METHOD FOR VARIABLE SELECTION IN REGRESSION MODELS

MARKO DIMOVSKI AND IRENA STOJKOVSKA

Abstract. A new type of regularization in least-square optimization for variable selection in regression models is proposed. Proposed regularization is suitable for regression models with equal or at least comparable regressors' influence. Consistency of the estimator of the regression parameter under suitable assumptions is shown. Numerical results demonstrate efficiency of the proposed regularization and its better performance compared to existing regularization methods.

1. Introduction

Variable selection in regression models means identifying the best subset among many variables to include in a model, which is maybe the hardest part of model building. Although, there is a theoretical advantage of eliminating irrelevant variables and, in some cases, even variables that contain some predictive information about the response variable, [8], it is also well known that omitting an important explanatory variable may produce severely biased parameter estimates and prediction results, [12].

Some variable selection methods have been developed to identify good (although not necessarily the best) subset models, with considerably less computing than it is required for all possible regressions. These methods are referred to as stepwise regression methods. The subset models are identified sequentially by adding or deleting, depending on the method (forward or backward stepwise selection respectively), the one variable that has the greatest impact on the residual sum of squares. But, neither forward selection nor backward elimination takes into account the effect that the addition or deletion of a variable can have, on the contributions of other variables to the model, [8].

Despite the stepwise regression methods, the Ordinary Least Square (OLS) method is one of the most widely used method for estimating the

2000 Mathematics Subject Classification. 62J05, 62J07, 93E24.

Key words and phrases. linear regression, regression models, least square method, regularization, penalty functions.

parameters in linear regression models. By minimizing the sum of residual square errors, the OLS method finds unbiased and consistent estimates. One of its drawbacks is overfitting the regression models, which can result in poor predictions. One way to overcome the overfitting is by adding an additional information to the model through regularization. One of the the well known regularization methods used in least-square optimization for variable selection in regression models are Ridge (or Tikhonov regularization), Least Absolute Shrinkage and Selection Operator or LASSO ([10]) and Elastic Net ([13]). Ridge regularization method improves the OLS estimates by continuous shrinkage of the regression coefficients, but LASSO regularization does both continuous shrinkage and automatic variable selection at the same time, and is successfully used when there are large number of predictors, such that the most relevant should be chosen. In comparison to these two regularization methods, Elastic Net, which is a combination of Ridge and LASSO, benefits from the both of them and gives better predictions. Another regularization method that can do both regression shrinkage and selection (like LASSO), but is resistent to outliers or heavy-tailed errors, is LED-lasso regularization method, proposed in [12].

When making a variable selection, LASSO and Elastic Net use L_1-norm for regularization, that can result in unwanted rejection of some predictors, especially when there is a group of predictors with very high pairwise correlation. One way to fix this drawback is by using L_∞-norm, as we have done in this work. When L_∞-norm is used, the predictors are less likely to be excluded from the model, so the model will have a chance to be built on the information from all its predictors and to fix the overfitting from OLS estimation. As it will be shown, the regularization method that we are proposing is the most suitable for models with equal or at least comparable regressors' influence. This kind of models can be found, for an example, in hydrology when exploring the dependence of river or lake water levels, on springs flows or underground water resources, [2]. We are also proposing a combination of L_2 and L_∞-norm for regularization, namely L_2-norm will help extreme values obtained by L_∞-norm to be avoided.

The paper is organized as following. In Section 2, OLS method and the existing regularization methods such as Ridge, LASSO and Elastic Net are presented. The new regularization method is presented in Section 3, where the consistency of the new estimator is established. A combination of L_2 and L_∞-norm for regularization, is also presented in Section 3. In Section 4, numerical comparative results obtained by testing OLS and five different regularization methods are presented. Conclusions are given in the last Section 5.

2. Preliminaries

We will consider the linear regression model given by it's matrix form

$$y = X\beta + \epsilon, \tag{1}$$

where:

- $y = (y_1, y_2, ..., y_n)^T$ is n-vector of dependent variables,
- $X = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]^T = [x_{ij}]_{n\times p}$ is $n \times p$-matrix of independent variables,
- $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$ is p-vector of associated regression coefficients, and
- $\epsilon = (\epsilon_1, \epsilon_2, ..., \epsilon_n)^T$ is n-vector which components are independent and identically distributed random errors with $E(\epsilon_i) = 0$ and $D(\epsilon_i) = \sigma^2$.

We assume that the matrix X is a non-stochastic matrix of known constants, with the full rank, [8].

The most commonly used method for estimating the unknown vector of parameters β in (1) is the Ordinary Least-Square (OLS) method, which minimizes the sum of residual square errors

$$RSS = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 .$$

In other words, the OLS parameter estimate $\hat{\beta}^{ols}$ is obtained by solving the unconstrained optimization problem i.e.

$$\hat{\beta}^{ols} = \underset{\beta\in\mathbb{R}^p}{\arg\min} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\}. \tag{2}$$

Despite of obtaining the estimates that are unbiased and consistent, the data analysts are often not satisfied with the OLS estimates. One of the reasons is the prediction accuracy, namely the OLS estimates often have low bias, but large variance. One way to overcome this deficiency and the overfitting in the regression model, is by introducing an additional information via the process of regularization. Using regularization penalties in model fitting is a very popular and successful approach in statistical modeling. Two popular regularization methods that are widely used are the Ridge (also known as Tikhonov regularization) and LASSO regularization, [10]. In the first one, the L_2-regularization penalty is used i.e. the constrained optimization problem

$$\min_{\beta\in\mathbb{R}^p} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \le t, \tag{3}$$

is considered, where $t \geq 0$ is the tuning parameter. Solving the constrained optimization problem (3) is equivalent to solving the unconstrained optimization problem obtained by adding penalty-like term in the objective function of (3) and introducing the Lagrangian multiplier $\lambda > 0$. So, the Ridge parameter estimate $\hat{\beta}^{ridge}$ is obtained as a solution of the equivalent unconstrained optimization problem i.e.

$$\hat{\beta}^{ridge} = \underset{\beta \in \mathbb{R}^{\prime}}{\arg\min} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}. \tag{4}$$

The tuning parameters t and λ control the amount of shrinkage that is applied to the estimates. Parameters t and λ have some kind of a reciprocal relationship, [10]. When t tends to infinity, than λ will be equal to zero, and the optimization problem (4) will be equivalent to the least-square optimization problem. If $t = 0$, then λ will tend to infinity, so the amount of shrinkage will be greater.

Ridge regression is a continuous process that shrinks coefficients and hence is a very stable process. However, it does not set any coefficients to 0 and hence does not give an easily interpretable model. For that reason, the LASSO regularization is proposed, [10]. This technique shrinks some of the regression coefficients and sets others to 0. LASSO uses L_1-norm penalty term and solves the constrained optimization problem

$$\min_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq t, \tag{5}$$

which is equivalent to finding an estimate $\hat{\beta}^{lasso}$ as a solution of the unconstrained optimization problem i.e.

$$\hat{\beta}^{lasso} = \underset{\beta \in \mathbb{R}^{\prime}}{\arg\min} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{6}$$

Due to the nature of the L_1 penalty, LASSO does both continuous shrinkage and automatic variable selection at the same time. Although LASSO has shown success in many situations, it has some limitations in some cases, especially when $p \ll n$ or when there is a group of variables among which the pairwise correlations are very high. In the second case, LASSO chooses only one variable from that group of variables, no matter which one.

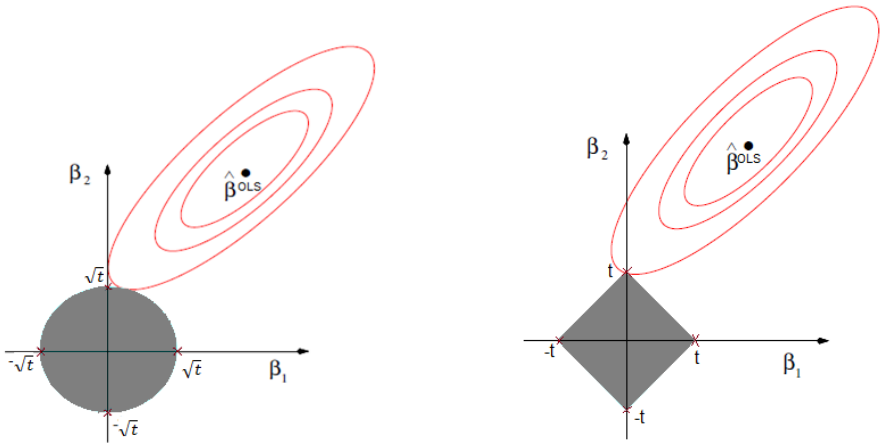
In order to provide some geometrical insight in these two regularization methods, Ridge and LASSO, let us first rewrite the OLS loss function

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2$$

in a form of a quadratic function

$$(\beta - \hat{\beta}^{ols})^T X^T X(\beta - \hat{\beta}^{ols}),$$

plus a constant, where $\hat{\beta}^{ols}$ is the OLS estimate. The elliptical contours of this function (for $p = 2$) are shown by the full curves in Figure 1, and they are centered at the OLS estimates. The constraint region for the Ridge regularization method is the disk $\beta_1^2 + \beta_2^2 \leq t$, while the constrained region for LASSO is the diamond $|\beta_1| + |\beta_2| \leq t$. Both methods find the first point where the elliptical contours hit the constraint region. Unlike the disk, the diamond has corners, and the LASSO solution might occur at a corner, corresponding to a zero coefficient, which is not a case when the constrained region is disk and zero solutions will rarely result.



a. Ridge regularization b. LASSO regularization

Figure 1: A geometric interpretation of Ridge and LASSO regularization methods for $p = 2$.

Elastic net regularization which is a combination of LASSO and Ridge, is introduced to obtain better prediction performance in a different situations, using the strengths of the both methods, [13]. The parameter estimator is found by solving the constrained optimization problem

$$\min_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\} \quad \text{s.t.} \quad \sum_{j=1}^{p} \beta_j^2 \leq t_1 \quad \text{and} \quad \sum_{j=1}^{p} |\beta_j| \leq t_2,$$

which is equivalent to finding an estimate $\hat{\beta}^{en}$ as a solution of the unconstrained optimization problem i.e.

$$\hat{\beta}^{en} = \underset{\beta\in\mathbb{R}^p}{\arg\min} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda_1 \sum_{j=1}^{p} \beta_j^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j| \right\}.$$

Elastic net regularization method overcomes the problems induced by having a small number of reliable observations for a large series of potential predictors i.e. $p \gg n$, found in genetic engineering or in chemometrics, [9]. Thanks to the added parameter $\lambda_2 > 0$, the total Elastic net penalty is strictly convex, and therefore Elastic net regression coefficients tend to be equal for highly correlated predictors, whereas LASSO assigns two different (biased) coefficients, [13].

3. New regularization in least-square optimization

3.1. **MASO regularization.** We are already familiar with the conclusion that we can not construct a regression model that will perfectly fit every type of data. Every method that we have already mentioned have it's own advantages and drawbacks depending on the underlying model and the nature of data. The idea for introducing a new type of regularization in the least-square optimization for variable selection in regression models is to improve the estimates from the least-square optimization and previous mentioned regularization methods, in the models with equal or at least comparable regressors' influence. As we mentioned before, the least-square optimization may result in overfitting the model, and the use of L_1-norm for regularization can overcome this problem, but may result in unwanted rejection of some predictors. So, one way to fix this last drawback is to introduce L_∞-norm in regularization term.

The regularization method that we propose uses L_∞ regularization penalty and finds estimates of the parameter vector in the linear regression model (1), by solving the constrained optimization problem

$$\min_{\beta\in\mathbb{R}^p} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\} \quad \text{subject to} \quad \max_{1\le j\le p} |\beta_j| \le t. \tag{7}$$

We name it Maximum Absolute Shrinkage Operator (MASO), because of the nature of L_∞-norm and its shrinkage property. The constrained problem (7) is equivalent to finding an estimate $\hat{\beta}^{maso}$ as a solution of the unconstrained optimization problem i.e.

$$\hat{\beta}^{maso} = \underset{\beta\in\mathbb{R}^p}{\arg\min} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \max_{1\le j\le p} |\beta_j| \right\}. \tag{8}$$

MASO regularization method is expected to overcome the problem when there are groups of variables with strong pairwise correlations in the regression model with equal or at least comparable regressors' influence, a situation when LASSO tends to select only one variable from each group.

A geometric interpretation of MASO regularization method in two - dimensional case i.e. for $p = 2$, Figure 2, provides us with some geometrical insight in this method. The constrained region is a square with vertices at points $(-t, -t)$, $(t, -t)$, (t, t) and $(-t, t)$. MASO solution is the first place where the contours of the OLS loss function touch the square, which sometimes may occur at a corner, corresponding to a coefficients with same absolute extreme value, which means that the corresponding predictors will have same influence in the regression model.

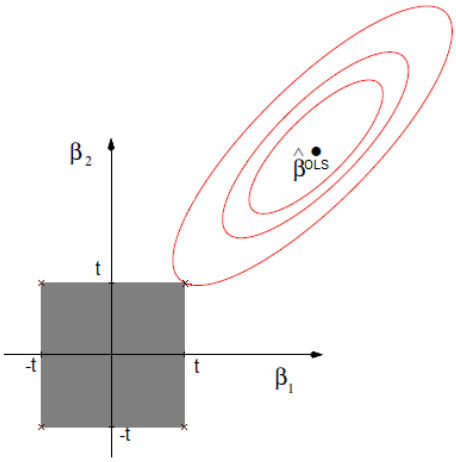


Figure 2: A geometric interpretation of MASO regularization method for $p = 2$

3.2. **Properties of MASO estimator.** To establish some properties of the MASO estimator $\hat{\beta}^{maso}$ defined by (8), we need to assume the following regularity conditions, [5], for the linear regression model (1):

Assumption 1. *There is a nonnegative definite matrix C such that*

$$C_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \to C, \tag{9}$$

and

$$\frac{1}{n} \max_{1 \le i \le n} \mathbf{x}_i^T \mathbf{x}_i \to 0. \tag{10}$$

As it is mentioned in [5], in practice, the covariates are usually scaled, so the diagonal elements of C_n (and of C) are all equal to 1.

Now, let us introduce a random function

$$Z_n(\phi) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\phi)^2 + \frac{\lambda_n}{n}\max_{1\le j\le p}|\phi_j|. \tag{11}$$

which is minimized at $\phi = \hat{\beta}_n$, a MASO estimator. The following theorem establishes the consistency of $\hat{\beta}_n$, under the condition that $\lambda_n = o(n)$.

Theorem 1. *Let the Assimption 1 holds. If C in (9) is nonsingular and $\frac{\lambda_n}{n}\to\lambda_0 \ge 0$, then*

$$\hat{\beta}_n \xrightarrow{p} \arg\min(Z),$$

where

$$Z(\phi) = (\phi-\beta)^T C(\phi-\beta) + \lambda_0 \max_{1\le j\le p}|\phi_j|,$$

and $\xrightarrow{p}$ denotes the convergence in probability. If $\lambda_n = o(n)$, then $\arg\min(Z) = \beta$, and the MASO estimator $\hat{\beta}_n$ is a consistent estimator for β.

Proof. We want to show that

$$\sup_{\phi\in K}|Z_n(\phi) - Z(\phi) - \sigma^2| \xrightarrow{p} 0 \tag{12}$$

for every compact set K and

$$\hat{\beta}_n = O_p(1). \tag{13}$$

Under (12) and (13), we will have

$$\hat{\beta}_n \xrightarrow{p} \arg\min(Z).$$

First, we will show the convexity of

$$f(\beta) = f(\beta_1, ..., \beta_p) = \max\{|\beta_1|, ..., |\beta_p|\}.$$

For $\lambda \in [0,1]$, and vectors $\beta = (\beta_1, ..., \beta_p), \gamma = (\gamma_1, ..., \gamma_p) \in \mathbb{R}^p$, we will denote by

$$|\beta_k| = \max\{|\beta_1|, ..., |\beta_p|\},\ |\gamma_s| = \max\{|\gamma_1|, ..., |\gamma_p|\}$$

and

$$|\lambda\beta_t + (1-\lambda)\gamma_t| = \max\{|\lambda\beta_1 + (1-\lambda)\gamma_1|, ..., |\lambda\beta_p + (1-\lambda)\gamma_p|\}.$$

Then,

$$\begin{aligned} & \lambda f(\beta) + (1-\lambda) f(\gamma) \\ = \ & \lambda \max\{|\beta_1|, ..., |\beta_p|\} + (1-\lambda)\max\{|\gamma_1|, ..., |\gamma_p|\} \\ = \ & \lambda|\beta_k| + (1-\lambda)|\gamma_s| \ge \lambda|\beta_t| + (1-\lambda)|\gamma_t| \\ = \ & |\lambda\beta_t| + |(1-\lambda)\gamma_t| \ge |\lambda\beta_t + (1-\lambda)\gamma_t| \end{aligned}$$

$$\begin{aligned} &= \max\{|\lambda\beta_1 + (1-\lambda)\gamma_1|, ..., |\lambda\beta_p + (1-\lambda)\gamma_p|\} \\ &= f(\lambda\beta + (1-\lambda)\gamma). \end{aligned}$$

So, we showed that

$$\lambda f(\beta) + (1-\lambda) f(\gamma) \geq f(\lambda\beta + (1-\lambda)\gamma),$$

for all $\lambda \in [0,1]$, and vectors $\beta, \gamma \in \mathbb{R}^p$, thus the function $f(\beta_1, ..., \beta_p) = \max\{|\beta_1|, ..., |\beta_p|\}$ is a convex function. The first term in (11) is a quadratic function, hence it is convex, so $Z_n(\phi)$ is a convex function as a sum of convex functions. Similarly, the function $Z(\phi) + \sigma^2$ is a convex function.

We will use the convexity lemma from [7] to prove (12). Besides the convexity of $Z_n(\phi)$ and $Z(\phi) + \sigma^2$, in order to apply the result of the convexity lemma from [7], we should show the pointwise convergence in probability of $Z_n(\phi)$ to $Z(\phi) + \sigma^2$, when $n \to \infty$. So, we have

$$\begin{aligned} Z_n(\phi) &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\phi)^2 + \frac{\lambda_n}{n}\max_{1\leq j\leq p}|\phi_j| \\ &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^T\beta - \mathbf{x}_i^T\phi + \epsilon_i)^2 + \frac{\lambda_n}{n}\max_{1\leq j\leq p}|\phi_j| \\ &= \frac{1}{n}\sum_{i=1}^{n}(\phi-\beta)^T\mathbf{x}_i\mathbf{x}_i^T(\phi-\beta) + \frac{2}{n}\sum_{i=1}^{n}\mathbf{x}_i^T(\beta-\phi)\epsilon_i \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 + \frac{\lambda_n}{n}\max_{1\leq j\leq p}|\phi_j|. \end{aligned}$$

Assumption 1 amd $\frac{\lambda_n}{n} \to \lambda_0$ imply that

$$\frac{1}{n}\sum_{i=1}^{n}(\phi-\beta)^T\mathbf{x}_i\mathbf{x}_i^T(\phi-\beta) + \frac{\lambda_n}{n}\max_{1\leq j\leq p}|\phi_j| \longrightarrow Z_(\phi).$$

We only need to show that

$$\frac{2}{n}\sum_{i=1}^{n}\mathbf{x}_i^T(\beta-\phi)\epsilon_i + \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 \xrightarrow{\ p\ } \sigma^2. \tag{14}$$

We know that $E(\epsilon_i) = 0$ and $D(\epsilon_i) = \sigma^2$, which implies that $E(\epsilon_i^2) = \sigma^2$. Hence, for $\delta > 0$,

$$P\{|\frac{1}{n}\sum_{i=1}^{n}(2\mathbf{x}_i^T(\beta-\phi)\epsilon_i + \epsilon_i^2) - \sigma^2| \geq \delta\} = P\{|U_n - EU_n| \geq \delta\},$$

where

$$U_n = \frac{1}{n}\sum_{i=1}^{n}(2\mathbf{x}_i^T(\beta-\phi)\epsilon_i + \epsilon_i^2)$$

is a random variable with the expectation $EU_n = \sigma^2$. From the Chebyshev inequality, we have that

$$P\{|\frac{1}{n}\sum_{i=1}^{n}(2\mathbf{x}_i^T(\beta-\phi)\epsilon_i+\epsilon_i^2)-\sigma^2|\geq\delta\}$$
$$\leq \frac{DU_n}{\delta^2}=\frac{1}{\delta^2 n^2}\sum_{i=1}^{n}d=\frac{d}{\delta^2 n}\longrightarrow 0,\quad \text{when}\quad n\to\infty,$$

where $d = DV_i$, $V_i = 2\mathbf{x}_i^T(\beta-\phi)\epsilon_i+\epsilon_i^2$, $i = 1, ..., n$. So, we proved (14). And because of the convexity lemma, [7], the convergence in probability given with (12) holds.

Let us notice that for the function (11) we have that

$$Z_n(\phi)\geq\frac{1}{n}\sum_{i=1}^{n}(y_i-\mathbf{x}_i^T\phi)^2=Z_n^{ols}(\phi), \tag{15}$$

for all ϕ, where $Z_n^{ols}(\phi)$ is the loss function for the OLS method. We know that $\hat{\beta}_n^{ols} = \arg\min(Z_n^{ols}(\phi)) = O_p(1)$ i.e. the OLS estimator is bounded with probability 1, because it is consistent (see [3]). This bound and (15) imply that (13) holds. Hence, from (12) and (13) the theorem is proven. □

Note that the model parameter vector β and the Lagrangian multiplier λ are indexed by n, since their values change with the growth of n.

We have already shown the consistency of the MASO estimator, but how fast does $\hat{\beta}_n$ converges to β? The following theorem indicates that we need $\lambda_n = O(\sqrt{n})$ for $\sqrt{n}$-consistency of the MASO estimator.

Theorem 2. *Let Asummption 1 holds. If C in (9) is nonsingular and $\frac{\lambda_n}{\sqrt{n}}\to\lambda_0\geq 0$, then*

$$\sqrt{n}(\hat{\beta}^n-\beta)\xrightarrow{d}\arg\min(V),$$

where $\xrightarrow{d}$ denotes the convergence in distribution, and

$$V(u)=-2u^TW+u^TCu+\lambda_0\max_{1\leq j\leq p}\{u_j\operatorname{sgn}(\beta_j)I(\beta_j\neq 0)+|u_j|I(\beta_j=0)\}$$

with $I(\cdot)$ as the indicator function, and W is a random vector with a multivariate normal distribution $W\sim\mathcal{N}(0,\sigma^2C)$.

Proof. Let us define a function

$$V_n(u)=\sum_{i=1}^{n}((\epsilon_i-\frac{u^T\mathbf{x}_i}{\sqrt{n}})^2-\epsilon_i^2)+\lambda_n\max_{1\leq j\leq p}\{|\beta_j+\frac{u_j}{\sqrt{n}}|-|\beta_j|\}. \tag{16}$$

For the function (11), we know that $\arg\min(Z_n(\phi)) = \hat{\beta}_n$, so

$$Z_n(\phi) \;=\; \frac{1}{n}\sum_{i=1}^{n}(y_i-\mathbf{x}_i^T\phi)^2+\frac{\lambda_n}{n}\max_{1\leq j\leq p}|\phi_j|$$

$$\begin{aligned} &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^T(\beta-\phi)+\epsilon_i)^2+\frac{\lambda_n}{n}\max_{1\le j\le p}|\phi_j| \\ &= \frac{1}{n}\sum_{i=1}^{n}(\epsilon_i-(\phi-\beta)^T\mathbf{x}_i)^2+\frac{\lambda_n}{n}\max_{1\le j\le p}|\phi_j| \\ &\ge \frac{1}{n}\sum_{i=1}^{n}(\epsilon_i-(\hat{\beta}_n-\beta)^T\mathbf{x}_i)^2+\frac{\lambda_n}{n}\max_{1\le j\le p}|[\hat{\beta}_n]_j|, \end{aligned}$$

for all ϕ, from where we have that

$$\begin{aligned} &\sum_{i=1}^{n}(\epsilon_i-(\phi-\beta)^T\mathbf{x}_i)^2+\lambda_n\max_{1\le j\le p}|\phi_j| \\ &\ge \sum_{i=1}^{n}(\epsilon_i-(\hat{\beta}_n-\beta)^T\mathbf{x}_i)^2+\lambda_n\max_{1\le j\le p}|[\hat{\beta}_n]_j|. \end{aligned}$$

We define a function $M(u)$ by

$$M(u)=\sum_{i=1}^{n}(\epsilon_i-\frac{u^T\mathbf{x}_i}{\sqrt{n}})^2+\lambda_n\max_{1\le j\le p}|\beta_j+\frac{u_j}{\sqrt{n}}|.$$

Let $u=\sqrt{n}(\phi-\beta)$, then we have

$$\begin{aligned} &M(\sqrt{n}(\phi-\beta)) \\ &= \sum_{i=1}^{n}(\epsilon_i-\frac{\sqrt{n}(\phi-\beta)^T\mathbf{x}_i}{\sqrt{n}})^2+\lambda_n\max_{1\le j\le p}|\beta_j+\frac{\sqrt{n}(\phi_j-\beta_j)}{\sqrt{n}}| \\ &= \sum_{i=1}^{n}(\epsilon_i-(\phi-\beta)^T\mathbf{x}_i)^2+\lambda_n\max_{1\le j\le p}|\phi_j| \\ &\ge \sum_{i=1}^{n}(\epsilon_i-(\hat{\beta}_n-\beta)^T\mathbf{x}_i)^2+\lambda_n\max_{1\le j\le p}|[\hat{\beta}_n]_j|, \end{aligned}$$

for all ϕ. So, if we put $\phi=u/\sqrt{n}+\beta$, then we will have that for all u the following inequality holds:

$$\begin{aligned} M(u) &= \sum_{i=1}^{n}(\epsilon_i-\frac{u^T\mathbf{x}_i}{\sqrt{n}})^2+\lambda_n\max_{1\le j\le p}|\beta_j+\frac{u_j}{\sqrt{n}}| \\ &\ge \sum_{i=1}^{n}(\epsilon_i-(\hat{\beta}_n-\beta)^T\mathbf{x}_i)^2+\lambda_n\max_{1\le j\le p}|[\hat{\beta}_n]_j|. \end{aligned} \tag{17}$$

We want to show that $\arg\min(V_n(u))=\sqrt{n}(\hat{\beta}_n-\beta)$, i.e. that for all u it holds that

$$V_n(u) = \sum_{i=1}^{n}((\epsilon_i-\frac{u^T\mathbf{x}_i}{\sqrt{n}})^2-\epsilon_i^2)+\lambda_n\max_{1\le j\le p}\{|\beta_j+\frac{u_j}{\sqrt{n}}|-|\beta_j|\}$$

$$\begin{aligned} &\geq \sum_{i=1}^{n}((\epsilon_i - \frac{\sqrt{n}(\hat{\beta}_n-\beta)^T\mathbf{x}_i}{\sqrt{n}})^2 - \epsilon_i^2) + \lambda_n \max_{1\leq j\leq p}\{|\beta_j + \frac{\sqrt{n}(\hat{\beta}_n-\beta)}{\sqrt{n}}| - |\beta_j|\} \\ &= \sum_{i=1}^{n}((\epsilon_i - (\hat{\beta}_n-\beta)^T\mathbf{x}_i)^2 - \epsilon_i^2) + \lambda_n \max_{1\leq j\leq p}\{|[\hat{\beta}_n]_j| - |\beta_j|\}. \end{aligned}$$

This inequality follows directly from (17), so we can conclude that $\arg\min(V_n(u)) = \sqrt{n}(\hat{\beta}_n - \beta)$ holds.

Next, we want to show that

$$\sum_{i=1}^{n}((\epsilon_i - \frac{u^T\mathbf{x}_i}{\sqrt{n}})^2 - \epsilon_i^2) \xrightarrow{d} -2u^TW + u^TCu. \tag{18}$$

Showing this convergence is equivalent to showing that

$$\sum_{i=1}^{n}(-2\frac{\epsilon_i u^T\mathbf{x}_i}{\sqrt{n}} + \frac{(u^T\mathbf{x}_i)^2}{n}) \xrightarrow{d} -2u^TW + u^TCu. \tag{19}$$

We will show that

$$\sum_{i=1}^{n} -2\frac{\epsilon_i u^T\mathbf{x}_i}{\sqrt{n}} \xrightarrow{d} -2u^TW \quad \text{and} \quad \sum_{i=1}^{n}\frac{(u^T\mathbf{x}_i)^2}{n} \xrightarrow{d} u^TCu, \tag{20}$$

which will imply that (19) and consequently (18) hold.

The term $u^T\mathbf{x}_i$ represent a scalar product, so $(u^T\mathbf{x}_i)^2 = (u^T\mathbf{x}_i)^T(u^T\mathbf{x}_i) = (u^T\mathbf{x}_i)(u^T\mathbf{x}_i)^T$ and

$$\begin{aligned} \sum_{i=1}^{n}\frac{(u^T\mathbf{x}_i)^2}{n} &= \sum_{i=1}^{n}\frac{(u^T\mathbf{x}_i)(u^T\mathbf{x}_i)^T}{n} = \sum_{i=1}^{n}\frac{u^T\mathbf{x}_i\mathbf{x}_i^Tu}{n} \\ &= u^T(\sum_{i=1}^{n}\frac{\mathbf{x}_i\mathbf{x}_i^T}{n})u \xrightarrow{d} u^TCu, \quad \text{as} \quad n\to\infty, \end{aligned} \tag{21}$$

due to Assumption 1. Now, because

$$\sum_{i=1}^{n} -2\frac{\epsilon_i u^T\mathbf{x}_i}{\sqrt{n}} = -2u^T\sum_{i=1}^{n}\frac{\epsilon_i\mathbf{x}_i}{\sqrt{n}},$$

and

$$\sum_{i=1}^{n}\frac{\epsilon_i\mathbf{x}_i}{\sqrt{n}} \xrightarrow{d} W, \tag{22}$$

due to the Central limit theorem for independent and identically distributed random variables, the Weak law of large numbers, and $W \sim \mathcal{N}(0, \sigma^2C)$, we have that the first part of (20) holds, which together with (21) imply (20).

At the end, we need to show that

$$\lambda_n \max_{1\leq j\leq p}\{|\beta_j + \frac{u_j}{\sqrt{n}}| - |\beta_j|\}$$

$$\longrightarrow \lambda_0 \max_{1\le j\le p}\{u_j \operatorname{sgn}(\beta_j) I(\beta_j \neq 0) + |u_j| I(\beta_j = 0)\}. \tag{23}$$

We need to consider two cases for β_j. If $\beta_j = 0$, then $I(\beta_j = 0) = 1$ and $I(\beta_j \neq 0) = 0$, so we will have that

$$\begin{aligned}&\lambda_n(|\beta_j + \frac{u_j}{\sqrt{n}}| - |\beta_j|) = \lambda_n|\frac{u_j}{\sqrt{n}}| \\ &\longrightarrow \lambda_0|u_j| = \lambda_0(u_j \operatorname{sgn}(\beta_j) I(\beta_j \neq 0) + |u_j| I(\beta_j = 0)).\end{aligned} \tag{24}$$

If $\beta_j \neq 0$, then $I(\beta_j = 0) = 0$ and $I(\beta_j \neq 0) = 1$. We will show that

$$\begin{aligned}&\lambda_n(|\beta_j + \frac{u_j}{\sqrt{n}}| - |\beta_j|) \\ &\longrightarrow \lambda_0(u_j \operatorname{sgn}(\beta_j)) = \lambda_0(u_j \operatorname{sgn}(\beta_j) I(\beta_j \neq 0) + |u_j| I(\beta_j = 0)).\end{aligned} \tag{25}$$

We have

$$\lim_{n\to\infty} \lambda_n(|\beta_j + \frac{u_j}{\sqrt{n}}| - |\beta_j|) = \lim_{n\to\infty} \frac{\lambda_n}{\sqrt{n}} \cdot \frac{|\beta_j + \frac{u_j}{\sqrt{n}}| - |\beta_j|}{\frac{1}{\sqrt{n}}}.$$

And because

$$\lim_{n\to\infty} \frac{|\beta_j + \frac{u_j}{\sqrt{n}}| - |\beta_j|}{\frac{1}{\sqrt{n}}} = \lim_{n\to\infty} \frac{\operatorname{sgn}(\beta_j + \frac{u_j}{\sqrt{n}})(-\frac{u_j}{2\sqrt{n^3}})}{-\frac{1}{2\sqrt{n^3}}} = u_j \operatorname{sgn}(\beta_j),$$

and $\frac{\lambda_n}{\sqrt{n}} \to \lambda_0$, we have

$$\lim_{n\to\infty} \lambda_n(|\beta_j + \frac{u_j}{\sqrt{n}}| - |\beta_j|) = \lambda_0(u_j \operatorname{sgn}(\beta_j)),$$

i.e. (25) holds. Now, (24) and (25) imply (23).

Finally, (18) and (23) imply

$$V_n(u) \xrightarrow{d} V(u).$$

Since, $V_n(u)$ is a convex function and the function $V(u)$ has a unique minimum, we have that

$$\arg\min(V_n(u)) = \sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} \arg\min(V(u)),$$

which proves the theorem. □

Note that the previous two consistency theorems are also valid for Ridge and LASSO estimators, and for Bridge estimators in general (the second one is valid for the Bridge estimators with parameter $\gamma \geq 1$), see [5].

3.3. **MASO-Ridge regularization.** The idea which led to combine the LASSO and Ridge regularization methods in one method that has the advantages of both types of regularization methods, is a good reason to consider a combination of MASO and Ridge regularization methods. Stability arising as a result of the tendency of Ridge regularization to obtain coefficients that tend to be equal when the predictors are highly correlated, is an enough reason to combine it with the newly introduced MASO regularization. So, the MASO-Ridge estimator can be obtained by solving the following constrained optimization problem:

$$\min_{\beta\in\mathbb{R}^p}\left\{\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2\right\}\quad\text{s.t.}\quad\sum_{j=1}^{p}\beta_j^2\le t_1\quad\text{and}\quad\max_{1\le j\le p}|\beta_j|\le t_2,$$

which is equivalent to finding a MASO-Ridge estimator $\hat{\beta}^{mr}$ by solving the unconstrained optimization problem:

$$\hat{\beta}^{mr}=\arg\min_{\beta\in\mathbb{R}^p}\left\{\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda_1\sum_{j=1}^{p}\beta_j^2+\lambda_2\max_{1\le j\le p}|\beta_j|\right\}.$$

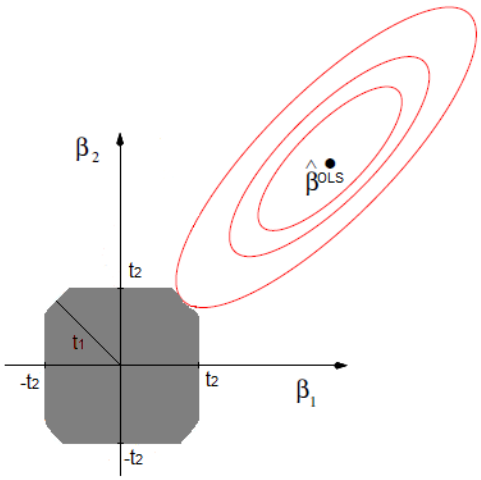


Figure 3: A geometric interpretation of MASO-Ridge regularization method for $p = 2$

To gain some geometrical insight at MASO-Ridge regularization method, in Figure 3, the constraint region and the elliptical contours of the OLS loss function, for $p = 2$, are shown. For $t_2 \le t_1 \le \sqrt{2}t_2$, the constraint region

has the form as shown on Figure 3, which means that the extreme values from MASO regularization can be avoided. If $t_1 \geq \sqrt{2}t_2$, the constraint region is the same as the square region from the MASO regularization. And if $t_1 \leq t_2$, the constraint region is the same as the one obtained in the Ridge regularization.

4. Simulation results

In this section we present the results from comparing the OLS estimates with the estimates obtained by the five regularization methods Ridge, LASSO, Elastic Net, MASO and MASO-Ridge, on two examples from [10]. To obtain the estimates, the underlying unconstrained optimization problems are solved using three different optimization methods: steepest descent method (SD), spectral projected gradient method (SPG) and stochastic approximation (SA). Each optimization method finds the next iterate by the iterative formula $\beta_{k+1} = \beta_k + \alpha_k p_k$, where β_k is the current iterate, and the search direction p_k is calculated by

$$p_k = -\nabla f(\beta_k),$$

for SD and SA method, and by

$$p_k = -\delta_k \nabla f(\beta_k),$$

for SPG method, with the spectral coefficient δ_k calculated as in [1]. Note that f is the loss function of the unconstrained optimization problem, and ∇f is it's gradient. The gradient is approximated by centered finite differences with step $h = 10^{-5}$.

In SD method, the step size α_k is calculated by the backtracking Armijo line search, with the coefficients proposed in [6]. In SPG method, the step size α_k is calculated by the nonmonotone line search with a safeguard quadratic interpolation, given in [1]. And, the step size α_k in SA method is calculated by

$$\alpha_k = \frac{a}{(k+1+A)^\gamma},$$

where $a = 0.01$, $A = 0$ and $\gamma = 0.602$.

Our goal is to estimate the parameter vector in the linear regression model $y = X\beta + \epsilon$ using simulated data and different estimation and optimization methods. We simulated $N = 50$ data sets of $n = 100$ observations, where the random errors ϵ_i are i.i.d. random variables with normal Gaussian distributions

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2),\ i = 1, 2, ..., n,$$

with $\sigma = 3$. The column vectors $X_j, j = 1, 2, ..., p$ of the matrix X of independent variables are chosen to have n-dimensional normal distributions

$$X_j \sim \mathcal{N}(0, C),\ j = 1, 2, ..., p,$$

where the covariance matrix $C = [c_{ij}]$ is such that

$$c_{ij} = \rho^{|i-j|}, i, j = 1, 2, ..., p,$$

with $\rho = 0.5$, [11]. The five-fold cross-validation is used to estimate the regularization parameter in each case, [4]. We choose $\{0, 0.01, 0.1, 1, 10, 100\}$ as a set of candidates for the optimal choice of the regularization parameter λ, [13].

Comparison of OLS method and regularization methods is based on the evaluation of Mean Square Errors (MSE) and Median Square Errors (MedianSE) defined by

$$MSE = \frac{1}{N}\sum_{k=1}^{N}(\hat{\beta}^k - \beta)^T C(\hat{\beta}^k - \beta)$$

and

$$MedianSE = Median\{(\hat{\beta}^k - \beta)^T C(\hat{\beta}^k - \beta),\ k = 1, 2, ..., N\},$$

respectively, where $\hat{\beta}^k$ is the kth estimate of the parameter β. The total number of zeros among all coefficients in 50 runs (TotalNZ) and the avarage number of zero coefficients at a single run (AverageNZ) have also been calculated, together with confidence intervals for every regression coefficient. All tests are conducted in MATLAB.

Example 1. [10] *In this example we are looking for the estimate of the parameter β in $y = X\beta + \epsilon$, where the true value of β is*

$$\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^T.$$

This is an example with most of the regression coefficients equal to zeros, so the expectations are that the best estimates will be obtained by the LASSO regularization or it's combination with the Ridge regularization i.e. the Elastic net regularization. We also expect that the MASO estimates will be an improvement of the OLS estimates, although they might not be as good as the LASSO estimates. The comparisons of different estimation methods in combination with different optimization methods are shown in Table 1.

As we can see from Table 1, the lowest MSE is obtained by the Elastic Net regularization with SPG optimization method, which is comparable to MSE from SA optimization method. Zero coefficients are only obtained by LASSO and Elastic Net regularization, and only one zero coefficient obtained by MASO regularization with SA optimization method. On this example MASO, and MASO-Ridge estimates manage to outperform (in MSE) only OLS estimates, as it was expected. Similar discussion can be conducted if we look at MedianSE, with LASSO and Elastic Net having the lowest MedianSE.

	MSE	MedianSE	TotalNZ	AverageNZ
OLS_SD	0.812069575	0.69996898	0	0
OLS_SPG	0.805519263	0.696021437	0	0
OLS_SA	0.812069575	0.69996898	0	0
Ridge_SD	0.826214041	0.689404581	0	0
Ridge_SPG	0.781524403	0.693105221	0	0
Ridge_SA	0.777243972	0.687549343	0	0
Lasso_SD	0.748250295	0.676306941	18	0.045
Lasso_SPG	0.733968123	0.67452177	17	0.0425
Lasso_SA	0.736761883	0.667729514	8	0.02
ElasticNet_SD	0.827978442	0.675987269	18	0.045
ElasticNet_SPG	0.728120685	0.668145177	15	0.0375
ElasticNet_SA	0.728137252	0.665618273	9	0.0225
Maso_SD	0.815829488	0.7177292	0	0
Maso_SPG	0.809944363	0.696021437	0	0
Maso_SA	0.804873102	0.689590035	1	0.0025
Maso-Ridge_SD	0.83072272	0.689404581	0	0
Maso-Ridge_SPG	0.788165377	0.693105221	0	0
Maso-Ridge_SA	0.783877838	0.687549343	0	0

Table 1: MSE, MedianSE, TotalNZ and AverageNZ (Example 1)

Example 2. [10] *In this example we are looking for the estimate of the parameter β in $y = X\beta + \epsilon$, where the true value of β is*

$$\beta = (0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85, 0.85)^T.$$

This is an example without any zero coefficient, and the same influence of the predictors in the regression model. So, we are expecting dominant performance of the newly introduced regularization methods, MASO and MASO-Ridge. Numerical results presented in Table 2 show that the lowest MSE is obtained from MASO-Ridge regularization with SA optimization method, which is comparable to the MSE from MASO regularization with SA optimization and to the MSE from MASO-Ridge regularization with SPG optimization. MedianSE acts in a similar way. Zero coefficients are again obtained by LASSO and Elastic Net regularization methods, and only one zero coefficient obtained by MASO-Ridge regularization with SPG optimization method. In this example, having a zero coefficient is undesirable.

We also present the 95% confidence intervals for every regression coefficients, in the both examples. They are shown in Table 3 and Table 4 respectively.

	MSE	MedianSE	TotalNZ	AverageNZ
OLS_SD	0.815670708	0.691576159	0	0
OLS_SPG	0.802919893	0.694733091	0	0
OLS_SA	0.798039516	0.68623488	0	0
Ridge_SD	0.873938864	0.692493117	0	0
Ridge_SPG	0.736557831	0.665858603	0	0
Ridge_SA	0.746653893	0.660054638	0	0
Lasso_SD	0.813968636	0.717586863	3	0.0075
Lasso_SPG	0.795488967	0.697667875	2	0.005
Lasso_SA	0.78975981	0.688818862	0	0
ElasticNet_SD	0.9325318	0.693088593	2	0.005
ElasticNet_SPG	0.738634681	0.668725098	1	0.0025
ElasticNet_SA	0.747933063	0.660564251	1	0.0025
Maso_SD	0.77011987	0.650267294	0	0
Maso_SPG	0.753124374	0.636702021	0	0
Maso_SA	0.72904457	0.641793174	0	0
Maso-Ridge_SD	0.772580122	0.632196319	0	0
Maso-Ridge_SPG	0.723636348	0.625733169	1	0.0025
Maso-Ridge_SA	0.703772749	0.628224325	0	0

Table 2: MSE, MedianSE, TotalNZ and AverageNZ (Example 2)

5. Conclusions

In this paper we introduced a new type of regularization in least-square optimization for model selection in regression. This regularization outperforms the OLS estimates by overcoming the overfitting, and has better performance compared to the existing regularization methods in models with equal or at least comparable regressors' influence. A consistency of the new estimator is established and a comparable simulation study is performed. Simulation results confirm our expectations.

Some of the future directions might be concerned with investigations in better selection of the regularization parameter, as well as with a combination of the proposed regularization with some other regularization methods.

Acknowledgements. This work is partially supported by Ss. Cyril and Methodius University of Skopje, Macedonia scientific research projects for 2014/2015 academic year.

	CI(β_1)	CI(β_2)	CI(β_3)	CI(β_4)	CI(β_5)	CI(β_6)	CI(β_7)	CI(β_8)
OLS_SD	2.904673325	1.34116174	−0.183945509	−0.060732595	1.936445879	−0.173430392	−0.184165667	−0.042141378
	3.094055281	1.58124345	0.04871818	0.129658465	2.126155681	0.079065838	0.044154892	0.173659807
OLS_SPG	2.903778132	1.339813516	−0.185076213	−0.063518883	1.933263336	−0.177816699	−0.187858573	−0.044624718
	3.093438458	1.579647133	0.048170289	0.126403686	2.121945693	0.073884106	0.042596491	0.170071256
OLS_SA	2.904673325	1.34116174	−0.183945509	−0.060732595	1.936445879	−0.173430392	−0.184165667	−0.042141378
	3.094055281	1.58124345	0.04871818	0.129658465	2.126155681	0.079065838	0.044154892	0.173659807
Ridge_SD	2.880975017	1.342005552	−0.171902206	−0.046078423	1.913649025	−0.153718401	−0.174557328	−0.037626742
	3.066748295	1.578705175	0.056700932	0.141935565	2.104861859	0.095736266	0.047819141	0.176946173
Ridge_SPG	2.875704652	1.338753655	−0.172208415	−0.051023356	1.903836263	−0.162078247	−0.184636894	−0.04163226
	3.062223668	1.572919022	0.055433539	0.134412156	2.088973284	0.083844505	0.040533001	0.168569796
Ridge_SA	2.875841592	1.340282797	−0.173129434	−0.048073422	1.900410041	−0.157554729	−0.18661723	−0.040495215
	3.061866121	1.57314242	0.053155929	0.137133255	2.084159105	0.087617551	0.037263209	0.168885421
Lasso_SD	2.879415893	1.317568954	−0.146604808	−0.049315855	1.896808814	−0.148105474	−0.164448813	−0.033178379
	3.06802623	1.551862218	0.063517105	0.122517719	2.0832832	0.080545251	0.04617547	0.168423799
Lasso_SPG	2.880161505	1.317170203	−0.149445085	−0.053711312	1.895293511	−0.152734187	−0.169553792	−0.037340609
	3.069172597	1.551356545	0.062836699	0.116898862	2.079965103	0.076534279	0.043203566	0.161711559
Lasso_SA	2.881814306	1.321272156	−0.158088542	−0.047431982	1.895411357	−0.153710478	−0.170973441	−0.039931093
	3.070999917	1.556081821	0.054338516	0.122780169	2.076978262	0.080812453	0.043076934	0.157373857
ElasticNet_SD	2.871969202	1.317676965	−0.141628539	−0.042393942	1.891178058	−0.13743237	−0.15412338	−0.030373767
	3.060286646	1.552795726	0.066383187	0.132166528	2.087044331	0.094623511	0.054659925	0.174371913
ElasticNet_SPG	2.870536114	1.319573068	−0.149577961	−0.047136691	1.886632235	−0.150313613	−0.167165614	−0.037677895
	3.059186438	1.552573861	0.062398665	0.122359813	2.069960534	0.07774883	0.043169812	0.159405972
ElasticNet_SA	2.870441951	1.320736462	−0.15391377	−0.042462013	1.884427092	−0.14781554	−0.169297192	−0.038090816
	3.059139346	1.554533534	0.057097793	0.126160252	2.065336392	0.083565579	0.041034875	0.156412319
Maso_SD	2.890267764	1.346269056	−0.184369773	−0.059023736	1.935621205	−0.172490076	−0.184208979	−0.041759449
	3.085533653	1.587273019	0.048292634	0.130347339	2.123205804	0.07999937	0.043876708	0.173825154
Maso_SPG	2.887446372	1.345039898	−0.185479367	−0.061990604	1.932504526	−0.176914515	−0.187858531	−0.044129162
	3.083865356	1.585960422	0.047705321	0.12709434	2.119135392	0.074669656	0.042339871	0.170390288
Maso_SA	2.889389864	1.345739308	−0.186103893	−0.058779999	1.928205183	−0.172330666	−0.189841386	−0.043011129
	3.084038512	1.585175171	0.045552565	0.130015511	2.11426702	0.078431491	0.039046889	0.170700385
Maso-Ridge_SD	2.873564191	1.345877223	−0.173638253	−0.045692979	1.915260252	−0.154667288	−0.174760328	−0.037718698
	3.062597192	1.58439337	0.05545394	0.141873018	2.106362717	0.095346544	0.047898179	0.177271245
Maso-Ridge_SPG	2.868752932	1.342847422	−0.174837316	−0.051245902	1.905791849	−0.163118797	−0.184633527	−0.041973012
	3.058444747	1.578898557	0.053790506	0.13392909	2.091938814	0.083341654	0.040812012	0.168751184
Maso-Ridge_SA	2.869257771	1.344011109	−0.175314572	−0.048829947	1.902404743	−0.158695669	−0.186550822	−0.040908459
	3.058414065	1.578576037	0.051760601	0.136458384	2.087680896	0.086874719	0.037554032	0.168969623

Table 3: 95% confidence intervals for regression coefficients (Example 1)

	CI(β_1)	CI(β_2)	CI(β_3)	CI(β_4)	CI(β_5)	CI(β_6)	CI(β_7)	CI(β_8)
OLS_SD	0.75552894	0.6908282	0.6675886	0.78825104	0.7894437	0.675581031	0.66837938	0.80757489
	0.944616355	0.930896001	0.899744782	0.978832819	0.979415307	0.927748234	0.896493081	1, 02378803
OLS_SPG	0.754279378	0.689765494	0.665723423	0.785634844	0.784669581	0.671647501	0.662739161	0.805299414
	0.943548895	0.929120684	0.898202579	0.975232902	0.97266477	0.922587019	0.892716554	1.019495303
OLS_SA	0.753999573	0.691060158	0.665926274	0.786413267	0.785046695	0.672526507	0.663266843	0.805787669
	0.94278736	0.929023548	0.896950835	0.975124526	0.971682554	0.922023155	0.892053137	1.019454228
Ridge_SD	0.75079707	0.699042838	0.684670992	0.793171527	0.780446043	0.686691908	0.674922308	0.795122603
	0.927369537	0.932316787	0.895344303	0.963291329	0.978482472	0.923260573	0.88847254	1.001688252
Ridge_SPG	0.73393182	0.688611018	0.670542588	0.776530513	0.772855386	0.671200478	0.658802145	0.789834776
	0.910462494	0.906881143	0.882591512	0.947815514	0.948482243	0.906035799	0.871836024	0.989202798
Ridge_SA	0.73767057	0.684269591	0.673287948	0.776549775	0.774741432	0.672743866	0.661025969	0.784220741
	0.916335847	0.905281715	0.886302127	0.94982468	0.950031005	0.906892555	0.877229501	0.988605125
Lasso_SD	0.744412401	0.687861439	0.674662627	0.779905502	0.782477883	0.672756014	0.66362184	0.806810634
	0.934962334	0.923979242	0.897710974	0.970663658	0.973370188	0.923583049	0.88795018	1.017469769
Lasso_SPG	0.744813817	0.687236865	0.673444206	0.77814531	0.778606533	0.668936032	0.657895075	0.80529724
	0.935069011	0.922771889	0.896942195	0.967124208	0.96730481	0.919018379	0.884204884	1.013663862
Lasso_SA	0.746481248	0.688021097	0.673370825	0.779834895	0.779222915	0.670630264	0.65859164	0.806029726
	0.935565466	0.922899367	0.895993393	0.968025573	0.96662266	0.919591621	0.88439139	1.014351769
ElasticNet_SD	0.750390989	0.699557096	0.690646235	0.790765786	0.783892333	0.688939489	0.679601051	0.793173406
	0.926649023	0.933320961	0.900405698	0.963062508	0.989245673	0.933313002	0.894615752	1.005626322
ElasticNet_SPG	0.733862232	0.687719086	0.675666098	0.774149269	0.772918137	0.668510343	0.660965569	0.79007778
	0.909817266	0.905409978	0.885528105	0.946739894	0.949463125	0.904330664	0.87485139	0.989201415
ElasticNet_SA	0.737150853	0.683111725	0.677328847	0.775855926	0.773932074	0.670461194	0.661558414	0.784890329
	0.915156465	0.903525919	0.887875877	0.949785229	0.950381662	0.905539848	0.877326413	0.988660504
Maso_SD	0.744894325	0.696036844	0.674561495	0.784063066	0.797581791	0.669453081	0.678124261	0.800183055
	0.917750163	0.928843852	0.89929276	0.964918656	0.981050346	0.907581701	0.891370602	1.004515905
Maso_SPG	0.749758547	0.693623545	0.669487826	0.785096204	0.78582129	0.674914205	0.665816245	0.807501524
	0.930377762	0.925839686	0.896907011	0.967292393	0.967541399	0.914069764	0.882800663	1.009803083
Maso_SA	0.740614626	0.703232718	0.666983577	0.781548144	0.79014039	0.665186593	0.670336133	0.79939901
	0.91192857	0.931984874	0.892016005	0.957040379	0.968612386	0.902351201	0.886753444	0.999726879
Maso-Ridge_SD	0.745000434	0.697182304	0.698462071	0.786338902	0.786564329	0.677744071	0.666962561	0.796936131
	0.911537142	0.929761048	0.913208033	0.949862216	0.969816185	0.907100943	0.873325452	0.994226799
Maso-Ridge_SPG	0.738791221	0.687395206	0.678316962	0.779617109	0.778483967	0.675948707	0.661806287	0.794720572
	0.912410212	0.904787594	0.89388658	0.95022548	0.951207765	0.904964063	0.87034886	0.995087161
Maso-Ridge_SA	0.729246327	0.698307054	0.677075524	0.77456638	0.782308064	0.674893186	0.660094111	0.792539368
	0.894454934	0.912627468	0.890801818	0.941013178	0.952796811	0.903525034	0.868508862	0.991885137

Table 4: 95% confidence intervals for regression coefficients (Example 2)

References

[1] E.N. Birgin, J. M. Martinez, M. Raydan, *Nonmonotone spectral projected gradient methods on convex sets*. SIAM J.Optim. Vol. 10, No. 4, (2014) pp. 1196-1211
[2] V. Bolshakov, *Regression-based Daugava River Flood Forecasting and Monitoring*. Snformation Technology and Management Science. Vol.16 (2013), 137-142
[3] R. Davidson, J. G. MacKinnon, *Econometric Theory and methods*, Oxford University Press, New York (2004)
[4] B. Efron, *The estimation of prediction error: Covariance penalties and crossvalidation*, J.Amer. Statis. Assoc., 99 (2004), 619-642
[5] K. Knight, W. Fu, *Asymptotics for Lasso-Type Estimators*, The Annals of Statistics Vol. 28 (5) (2000) 1356-1378
[6] J. Nocedal, S.J. Wright, *Numerical Optimization*, 2nd edition. Springer (2006
[7] D. Pollard, *Asymptotics for least absolute deviation regression estimators*, Econometric Theory Vol.7 (1991) 186-199
[8] J. O. Rawlings, S. G. Pentula, D. A. Dickey, *Applied regression analysis: a reseaech tool, 2nd edition*, Springer-Verlag, New York (1998)
[9] I. Savin, *A Comparative Study of the Lasso-type and Heuristic Model Selection Methods*, Journal of Economics and Statistics Vol. 233 (4) (2013) 526-549
[10] R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society, Series B Vol.58 (1996) 267-288
[11] R. Tibshirani, *The Lasso method for variable selection in the Cox model*, Statistics in medicine, Vol.16 (1997) 385-395
[12] H. Wang, G. Li, G. Jiang, *Robust regression shrinkage and consistent variable selection through the LAD-Lasso*, J Bus. Econ. Stat., Vol.25(3) (2007) 347-355
[13] H. Zou, T. Hastie, *Regularization and variable selection via the elastic net*, J.R.Statist, Soc.B.Vol.67, Part 2 (2005) 301-320

Department of Mathematics, Faculty of Natural Sciences and Mathematics, Ss. Cyril and Methodius University, Arhimedova 3, 1000 Skopje, Macedonia.

E-mail address: mdimovski16@gmail.com

Department of Mathematics, Faculty of Natural Sciences and Mathematics, Ss. Cyril and Methodius University, Arhimedova 3, 1000 Skopje, Macedonia.

E-mail address: irenatra@pmf.ukim.mk