

Regularization and variable selection with triple shrinkage in linear regression: a generalization of lasso

Murat Genç & M. Revan Özkale

To cite this article: Murat Genç & M. Revan Özkale (2024) Regularization and variable selection with triple shrinkage in linear regression: a generalization of lasso, Communications in Statistics - Simulation and Computation, 53:11, 5242-5264, DOI: [10.1080/03610918.2023.2173780](https://doi.org/10.1080/03610918.2023.2173780)

To link to this article: <https://doi.org/10.1080/03610918.2023.2173780>



Published online: 06 Feb 2023.



Submit your article to this journal [↗](#)



Article views: 378



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 2 View citing articles [↗](#)



Regularization and variable selection with triple shrinkage in linear regression: a generalization of lasso

Murat Genç^a and M. Revan Özkale^b

^aDepartment of Management Information Systems, Faculty of Economics and Administrative Sciences, Tarsus University, Mersin, Turkey; ^bDepartment of Statistics, Çukurova University, Faculty of Science and Letters, Adana, Turkey

ABSTRACT

We propose a new shrinkage and variable selection method in linear regression, which is based on triple shrinkage on the regression coefficients. The new estimation method contains the ridge, lasso and elastic net as special cases. The term based on the shrunken estimator in the new method can provide estimates with a smaller length depending on the size of a new tuning parameter compared to the elastic net, maintaining the variable selection feature in the case of multicollinearity. The new estimator has the property of the grouping effect similar to that of the elastic net. The well-known coordinate descent algorithm is used to compute the coefficient path of the new estimator, efficiently. We conduct real data analysis and simulation studies to compare the new estimator with several methods including the lasso and elastic net.

ARTICLE HISTORY

Received 20 August 2022
Accepted 23 January 2023

KEYWORDS

Coordinate descent algorithm; Elastic net; Grouping property; Lasso; Shrinkage; Variable selection

1. Introduction

We use statistical models to simplify and interpret complex phenomena with uncertainty based on the information in the observed data set. A statistical model describes the data, determines the functional expression that best represents the data set according to a certain criterion and depicts the relationship among the explanatory variables.

Linear regression models are special types of statistical models used to analyze the relationship between the response and the explanatory variables. Two important concepts in linear regression modeling are the predictive accuracy and the interpretability of the model. These concepts are evaluated with respect to the model coefficients. Predictive accuracy is a measure of the closeness between the estimated response values and the actual response values. An estimator yielding low variance is preferable in that it exhibits high prediction accuracy. The model interpretability is related to the modeling of the explanatory variables that have the strongest effects on the response variable. When a linear regression model contains explanatory variables that best describe the response variable, the model has a high predictive performance which is a desirable situation.

Although the ordinary least squares (OLS) is the well-known estimation method in linear regression, it has some drawbacks. If there is multicollinearity which is an approximate linear dependency among the explanatory variables in the data set, the estimates of the OLS coefficients show poor predictive performance. Another weakness of the OLS method is that no model

coefficient estimate is zero. Hence, the OLS can not discard variables from the model that are not related to the response variable. This leads to poor results in terms of model interpretability.

Various methods have been proposed to improve the OLS estimates. One of which is based on the penalization of regression coefficients. In penalized regression methods, the regression coefficients are estimated by using some penalties on the coefficients. In these methods, it is generally aimed to decrease the variance of the coefficients by introducing some bias to the regression coefficients. The subset selection method corresponds to ℓ_0 -norm penalization. In this method, explanatory variables that have strong effects on the response variable are retained in the model. For this reason, the model interpretability of this method is high. However, the estimates obtained in this model are quite unstable, i.e., a small change in the data set leads to a completely different model. The ridge regression proposed by Hoerl and Kennard (1970) corresponds to ℓ_2 -norm penalization. This method shrinks the coefficients in a continuous process according to the tuning parameter contained in the model. Thus, the method yields coefficients with a smaller variance than the subset selection method. However, no model coefficient estimated from this method is equal to zero. Therefore, ridge regression gives poor results in terms of model interpretability. The lasso (least absolute shrinkage and selection operator) proposed by Tibshirani (1996) is based on ℓ_1 -norm penalization. The lasso shrinks the coefficients toward zero as in the ridge regression setting. However, unlike ridge regression, some of the coefficient estimates are set exactly zero due to the nature of ℓ_1 -norm. This method, therefore, performs automatic variable selection besides shrinking the model coefficients. The lasso yields very good results in terms of prediction accuracy and model interpretability, if the true model is sparse (Tibshirani 1996); however, the lasso does not yield good estimates if multicollinearity exists and it does not have a grouping property. After the lasso is given by Tibshirani (1996), new adaptations of the lasso have been proposed to improve deficiencies or weaknesses arising from the diversity of data. The elastic net (Zou and Hastie 2005) was proposed to deal with correlated group variables, fused lasso (Tibshirani et al. 2005) was proposed to improve the lasso's inability to ordering features, group lasso (Yuan and Lin 2006) to solve groups of related explanatory variables such as dummy variables or ANOVA models, adaptive lasso (Zou 2006) to solve oracle property, twin adaptive lasso (Lee, Shi, and Gao 2021) to overcome the deficiency of adaptive lasso not eliminating all cointegrating variables with the zero regression coefficients, fuzzy adaptive lasso (Kong 2022) for fuzzy linear regression with crisp inputs and fuzzy outputs, Pliable lasso (Tibshirani and Friedman 2020) as a generalization of the lasso that allows the model coefficients to vary as a function of a general set of some prespecified modifying variables, two stage control function method (Xu, Li, and Zhang 2020) for econometrics when endogeneity exists in high-dimensional sparse situation, regularized structural equation modeling (Jacobucci, Grimm, and McArdle 2016) which is an extension usage of the lasso and ridge estimators to structural equation models, joint lasso (Dondelinger, Mukherjee, and Alzheimer's Disease Neuroimaging Initiative 2020) which jointly estimates the regression coefficients that induces global sparsity and encourages similarity between subgroup-specific coefficients (shares similarities with both group lasso and fused lasso), etc. Our contribution to the literature will be in this regard.

In this study, we propose a new penalization method and examine its characteristics. For the sake of brevity, we call the new estimator the GO¹ estimator. The GO estimator is based on the triple shrinkage of the coefficients. It enjoys the variable selection property of the lasso like the elastic net; however, it can produce estimates close to the shrunken estimates rather than the zero estimates when estimates are compared to the elastic net as a result of the nature of the third term in the penalization function. It also has a type of grouping property as we will explain the details in the following sections.

¹The word GO both shows the abbreviations of the authors' names and also has the meaning of continuing to be in a particular state which carries the idea of ridge, lasso and elastic net in a further way.

We organize the paper as follows. In [Sec. 2](#), we mention the well-known penalized methods such as the ridge, lasso and elastic net. In [Sec. 3](#), we introduce the GO estimator and examine its properties, besides, we propose a coordinate descent algorithm to compute the solution path and give a theorem about the alternative grouping property of the GO estimator. We compare the OLS, ridge, lasso, elastic net and GO estimators on a real data set in [Sec. 4](#) and on simulation studies in [Sec. 5](#). We explain the details of the modification of the GO estimator to the high-dimensional data in [Sec. 6](#). Finally, we end the study with the conclusions in [Sec. 7](#).

2. Penalized regression methods

We consider the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (1)$$

where $\mathbf{y} = (y_1, y_2, \dots, y_n)^\top$ is the $n \times 1$ response vector, $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p)$ is the $n \times p$ design matrix of $n \times 1$ explanatory variables $\mathbf{x}_j = (x_{1j}, x_{2j}, \dots, x_{nj})^\top$, $j = 1, 2, \dots, p$ with $n > p$, $\boldsymbol{\beta}$ is the $p \times 1$ coefficient vector of the model and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)^\top$ is the $n \times 1$ error vector. The error terms are assumed to have mean zero, constant variance σ^2 and uncorrelated.

In this study, it is assumed that the response variable is centered to have mean zero, $n^{-1} \sum_{i=1}^n y_i = 0$, and the explanatory variables are standardized to be centered to have mean zero with unit variance: $n^{-1} \sum_{i=1}^n x_{ij} = 0$ and $n^{-1} \sum_{i=1}^n x_{ij}^2 = 1$ for $j = 1, 2, \dots, p$. The coefficient vector $\boldsymbol{\beta}$ is generally estimated by minimizing the error sum of squares which results in the OLS estimator: $\hat{\boldsymbol{\beta}}^{ls} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ given that the design matrix is full column rank.

Penalized regression methods, which have been widely used (Hastie, Tibshirani, and Friedman 2009), have been proposed to deal with multicollinearity in the data or in the case of high-dimensional data, where the number of parameters is much larger than the number of observations. In penalized regression, the objective function is defined as

$$Q(\lambda, \boldsymbol{\beta})^2 = \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda P(\boldsymbol{\beta}) \quad (2)$$

²where $P(\cdot)$ is the penalization function and $\lambda > 0$ is known as the tuning parameter.

Some of the well-known penalized regression methods are the ridge regression, lasso and elastic net methods. Parameter estimates of ridge regression are obtained by minimizing [Equation \(2\)](#) with the penalization function $P(\boldsymbol{\beta}) = \frac{1}{2} \|\boldsymbol{\beta}\|_2^2$ which leads to

$$\hat{\boldsymbol{\beta}}^r = \left(\frac{1}{n} \mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I} \right)^{-1} \frac{1}{n} \mathbf{X}^\top \mathbf{y}$$

where $\lambda > 0$ determines the amount of shrinkage.

The disadvantage of ridge regression is due to the choice of λ . As λ goes to infinity from zero, $\hat{\boldsymbol{\beta}}^r$ goes from $\mathbf{0}$ to $\hat{\boldsymbol{\beta}}^{ls}$ where $\mathbf{0}$ is biased stable estimator of $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}^{ls}$ is unbiased unstable estimator of $\boldsymbol{\beta}$. Then as λ goes to infinity from zero, the expected distance between $\hat{\boldsymbol{\beta}}^r$ and $\boldsymbol{\beta}$ decreases. However, there is no optimum λ value that produces such a point estimate. These idea let Özkale and Kaç İranlar (2007) to define a new estimator. Their estimator minimizes [Equation \(2\)](#) with penalization function $P(\boldsymbol{\beta}) = \frac{1}{2} \|\boldsymbol{\beta} - d\hat{\boldsymbol{\beta}}^{ls}\|_2^2$ which was called OK estimator by Gruber (2012)³.

²In some descriptions, the factor $\frac{1}{2n}$ in [Equation \(2\)](#) can be replaced by $\frac{1}{2}$ or 1. According to the Hastie, Tibshirani, and Wainwright (2015), this case makes no difference in the objective function and corresponds to a reparametrization of λ in [Equation \(2\)](#). Advantage of the form in [Equation \(2\)](#) is stated by Hastie, Tibshirani, and Wainwright (2015) as that this kind of standardization makes λ values comparable for different sample sizes.

³The OK abbreviation denotes the first letters of the names of the authors.

$d\hat{\beta}^{ls}$ is chosen as the shrunken estimator of Mayer and Willke (1973) where $0 \leq d < 1$ and shows the prior information on β . The OK estimator has a closed-form solution as

$$\hat{\beta}(\lambda, d) = \left(\frac{1}{n} \mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_p \right)^{-1} \left(\frac{1}{n} \mathbf{X}^\top \mathbf{y} + \lambda d \hat{\beta}^{ls} \right). \quad (3)$$

The penalized regression methods can also make simultaneously shrinkage and variable selection according to the special case of the penalization function. The lasso, which is based on the objective function (2) with the penalization function $P(\beta) = \|\beta\|_1$ where $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$ is the ℓ_1 -norm of the coefficients was proposed by Tibshirani (1996) so as to obtain sparse solutions.

Although the lasso is a useful method for sparse modeling, the problems that occur with the lasso are that ridge regression gives better results in terms of mean squared error than the lasso for $n > p$ in the case of multicollinearity (Tibshirani 1996), the lasso models as many variables as the number of observations when the data set is high-dimensional (Tibshirani 1996) and the lasso models only one of the variables in the highly correlated group variables (Zou and Hastie 2005). In such cases, alternative methods may be used instead of the lasso, based on the generalization of the penalty function of the lasso.

Zou and Hastie (2005) proposed the elastic net as an alternative to the lasso. The penalization function of the elastic net is a combination of ℓ_1 and ℓ_2 -norm penalty terms. Given $\lambda_1 > 0$ and $\lambda_2 > 0$, elastic net parameter estimates are obtained by minimizing the objective function

$$Q(\lambda_1, \lambda_2, \beta) = \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda_1 \|\beta\|_1 + \frac{\lambda_2}{2} \|\beta\|_2^2.$$

An appropriate reparametrization is based on $\lambda = \lambda_1 + \lambda_2$ and $\alpha = \lambda_1 / (\lambda_1 + \lambda_2)$. This entails the objective function (2) with penalization function $P(\beta) = \alpha \|\beta\|_1 + \frac{1-\alpha}{2} \|\beta\|_2^2$ where $0 \leq \alpha \leq 1$ ⁴. The advantage of the elastic net over the lasso is that it has the grouping effect where all the group variables⁵ are modeled by assigning coefficients different from zero.

The ridge, lasso, elastic net and OK methods yield biased parameter estimates with relatively good prediction performance. Although the ridge and lasso are imposed by a single shrinkage, the elastic net and OK are incurred a double shrinkage. This double shrinkage in elastic net is applied by first finding the ridge estimates for each fixed λ_2 and then solving the lasso along the λ_1 values. The double shrinkage does not provide a significant reduction in variance in the elastic net. Besides, the bias in elastic net estimates is more pronounced due to the double shrinkage when compared with the lasso and ridge methods (Zou and Hastie 2005).

3. The GO estimator

Our aim is now to find an estimator which combines the advantages of the elastic net and shrunken estimator given by Mayer and Willke (1973) as $d\hat{\beta}^{ls}$, $0 \leq d < 1$. This new estimator will be proposed in such a way that its length is closer to the true parameter vector than the OLS estimator and carries the variable selection and grouping properties of the elastic net.

Let $0 \leq d < 1$, $\lambda_1 > 0$ and $\lambda_2 > 0$ be fixed tuning parameters. The proposed estimator is based on minimizing the objective function⁶

⁴This penalty is originally called as naive elastic net. Zou and Hastie (2005) used rescaled version of this penalty and called as elastic net. But Friedman, Hastie, and Tibshirani (2010) drops this distinction as we do in this paper.

⁵If the coefficients of highly correlated variables are approximately equal in absolute value, these variables are called group variables.

⁶Arashi, Asar, and Yüzbaşı (2021) proposed an estimator in a similar form to Equation (4). Their approach is based on fixing two of the three tuning parameters and giving an approximate solution for the estimator. In this manuscript, we follow a more general approach by considering flexible tuning parameters and an exact solution. Also, we investigate different properties of the estimator than the properties of the estimator in Arashi, Asar, and Yüzbaşı (2021).

$$Q(\boldsymbol{\beta}; d, \lambda_1, \lambda_2) = \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \frac{\lambda_2}{2} \|\boldsymbol{\beta} - d\hat{\boldsymbol{\beta}}^{ls}\|_2^2. \quad (4)$$

For known values of d , the function

$$\lambda_1 \|\boldsymbol{\beta}\|_1 + \frac{\lambda_2}{2} \|\boldsymbol{\beta} - d\hat{\boldsymbol{\beta}}^{ls}\|_2^2 \quad (5)$$

is the penalization function of the new estimator. For usual $n > p$ cases, if there is multicollinearity in the data set, it has been empirically observed that the prediction performance of the lasso is dominated by ridge regression (Tibshirani 1996). Therefore, with the new estimator, it is possible to reinforce further the prediction performance of the lasso.

Defining $\lambda = \lambda_1 + \lambda_2$ and $\alpha = \lambda_1/(\lambda_1 + \lambda_2)$, then the function (4) becomes

$$Q(\boldsymbol{\beta}; d, \lambda, \alpha) = \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \left\{ \alpha \|\boldsymbol{\beta}\|_1 + \frac{1-\alpha}{2} \|\boldsymbol{\beta} - d\hat{\boldsymbol{\beta}}^{ls}\|_2^2 \right\}, \quad (6)$$

where $\lambda > 0$, $0 \leq \alpha \leq 1$ and $0 \leq d < 1$. We prefer to minimize the function in the form of (6) for convenience.

We also note that the minimization of the function (6) is equivalent to the problem in the constrained form

$$\arg\min_{\boldsymbol{\beta}} \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \quad \text{subject to} \quad \alpha \|\boldsymbol{\beta}\|_1 + \frac{1-\alpha}{2} \|\boldsymbol{\beta} - d\hat{\boldsymbol{\beta}}^{ls}\|_2^2 \leq t \quad \text{for some } t \quad (7)$$

with $0 \leq \alpha \leq 1$.

As the consequence of using the penalization function in Equation (6), we obtain an estimator which is a generalization of the OLS, ridge, OK, lasso and elastic net estimators for special cases of the tuning parameters: lasso for $\alpha = 1$, OK for $\alpha = 0$, elastic net for $d = 0$, ridge for $\alpha = d = 0$ and OLS for $\lambda = 0$. We call the new estimator obtained by minimizing $Q(\boldsymbol{\beta}; d, \lambda, \alpha)$ as the GO estimator. The GO estimator has the following properties:

- (i) For fixed λ and d values, since the GO solution reduces to the OK solution when $\alpha = 0$ and the lasso solution when $\alpha = 1$, it traces a curve path through the parameter space from the OK estimator to the lasso as α goes from 0 to 1. That is, the GO solutions oscillate between the OK and lasso solutions for $0 \leq \alpha \leq 1$.
- (ii) It is able to select sub-models as in the lasso case due to the ℓ_1 -norm term in the objective function. Hence, it does variable selection as the lasso and elastic net.
- (iii) For any $\alpha < 1$, $\lambda > 0$ and $0 \leq d < 1$, the problem in Equation (6) is strictly convex. Therefore, the GO estimator is the unique solution irrespective of the correlations and duplications in \mathbf{x}_j .
- (iv) The ℓ_2 -norm term included in the objective function has the potential to maintain coefficient estimates with a length closer to $\boldsymbol{\beta}$ than the OLS depending on the parameter d in the case of multicollinearity.
- (v) When α and λ^* are fixed, the elastic net has a shorter length than the lasso along the λ values which are longer than or equal to a λ^* value which is a specific value of the tuning parameter values used in computing the solutions. However, when α, d and λ^* are fixed, the GO solution is closer to $d\hat{\boldsymbol{\beta}}^{ls}$ than the lasso solutions along the λ values which are longer than or equal to a λ^* value which is a specific value of the tuning parameter values used in computing the solutions. This case supports efficiency over the elastic net when the length of the estimator will be far away from the true parameter in the presence of multicollinearity.

The objective function given by Equation (4) depends on the parameters d, λ_1, λ_2 while the reparametrized form of the objective function given by Equation (6) depends on the parameters d, α, λ . While both forms can be used, we fit the models by using the reparametrized form because α and d are bounded in the interval $[0, 1]$ which is in a similar form to the elastic net described in Hastie, Tibshirani, and Wainwright (2015). Therefore we select a grid of d values in the interval $[0, 1)$ and, for each d value, select a grid of α values. Then for each pair of $\{d, \alpha\}$ we select a grid of λ values and find the best value of λ by cross-validation. Finally, we fit the model by using the optimal values of the tuning parameters d, α and λ . This tuning parameter selection method can be seen as an extension of the method described in Hastie, Tibshirani, and Wainwright (2015) to the new estimator by the new tuning parameter d .

3.1. ℓ_1 adjustment for the GO estimator

The relationship between the GO and lasso estimators can be shown by making use of augmented matrices. We assume λ_1 and λ_2 are known. Given the design matrix \mathbf{X} and response vector \mathbf{y} , we define $(n+p) \times p$ matrix \mathbf{X}^* and $(n+p) \times 1$ vector \mathbf{y}^* as follows:

$$\mathbf{X}^* = \begin{pmatrix} \mathbf{X} \\ \sqrt{n\lambda_2}\mathbf{I} \end{pmatrix} \text{ and } \mathbf{y}^* = \begin{pmatrix} \mathbf{y} \\ \sqrt{n\lambda_2}d\hat{\boldsymbol{\beta}}^{ls} \end{pmatrix}. \quad (8)$$

Then, it is straightforward to show that

$$\begin{aligned} \frac{1}{2n} \|\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}\|_2^2 &= \frac{1}{2n} \left(\begin{pmatrix} \mathbf{y} \\ \sqrt{n\lambda_2}d\hat{\boldsymbol{\beta}}^{ls} \end{pmatrix} - \begin{pmatrix} \mathbf{X} \\ \sqrt{n\lambda_2}\mathbf{I} \end{pmatrix} \boldsymbol{\beta} \right)^\top \left(\begin{pmatrix} \mathbf{y} \\ \sqrt{n\lambda_2}d\hat{\boldsymbol{\beta}}^{ls} \end{pmatrix} - \begin{pmatrix} \mathbf{X} \\ \sqrt{n\lambda_2}\mathbf{I} \end{pmatrix} \boldsymbol{\beta} \right) \\ &= \frac{1}{2n} \left[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + n\lambda_2 (d\hat{\boldsymbol{\beta}}^{ls} - \boldsymbol{\beta})^\top (d\hat{\boldsymbol{\beta}}^{ls} - \boldsymbol{\beta}) \right] \\ &= \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \frac{\lambda_2}{2} \|\boldsymbol{\beta} - d\hat{\boldsymbol{\beta}}^{ls}\|_2^2. \end{aligned}$$

Therefore, under the equalities in (8), we obtain

$$\frac{1}{2n} \|\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 = \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \frac{\lambda_2}{2} \|\boldsymbol{\beta} - d\hat{\boldsymbol{\beta}}^{ls}\|_2^2$$

which indicates that the GO estimator is equivalent to the lasso estimator based on augmented design matrix \mathbf{X}^* and response vector \mathbf{y}^* . Hence, the GO estimator enjoys the computational advantages of the lasso as the elastic net does.

3.2. Coordinate descent algorithm for the GO estimator

The penalization function in Equation (5) is separable in its components. Accordingly, the coordinate descent algorithm can be used to compute the estimates of the GO estimator. We write the objective function in terms of $\beta_j, j = 1, 2, \dots, p$:

$$f(\beta_j) = \frac{1}{2n} \sum_{i=1}^n (r_{ij} - x_{ij}\beta_j)^2 + \lambda_1 \sum_{j=1}^p |\beta_j| + \frac{\lambda_2}{2} \sum_{j=1}^p (\beta_j - d\hat{\beta}_j^{ls})^2$$

where $r_{ij} = y_i - \sum_{k \neq j} x_{ik}\beta_k$ is called the partial residual of the variable \mathbf{x}_j . The sub-differential of the function $f(\beta_j)$ is

$$\partial f(\beta_j) = \left(\frac{1}{n} \sum_{i=1}^n x_{ij}^2 + \lambda_2 \right) \beta_j - \frac{1}{n} \sum_{i=1}^n x_{ij} r_{ij} - \lambda_2 d \hat{\beta}_j^{ls} + \lambda_1 s_j, \quad j = 1, 2, \dots, p$$

where

$$s_j = \begin{cases} -1, & x < 0 \\ [-1, 1], & x = 0. \\ 1, & x > 0 \end{cases}$$

Hence, by equating $\partial f(\beta_j)$ to zero, we get the estimates of the new estimator in an iterative manner as

$$\hat{\beta}_j \leftarrow \frac{S\left(\frac{1}{n} \sum_{i=1}^n x_{ij} \hat{r}_{ij} + \lambda_2 d \hat{\beta}_j^{ls}, \lambda_1\right)}{\frac{1}{n} \sum_{i=1}^n x_{ij}^2 + \lambda_2}, j = 1, 2, \dots, p$$

where $S(\cdot)$ is the soft-thresholding operator defined by Donoho and Johnstone (1994). When the explanatory variables have mean zero and standard deviation 1, as a special case, the iteration steps for the solution are

$$\hat{\beta}_j \leftarrow \frac{S\left(\frac{1}{n} \sum_{i=1}^n x_{ij} \hat{r}_{ij} + \lambda_2 d \hat{\beta}_j^{ls}, \lambda_1\right)}{1 + \lambda_2}.$$

We, now, examine the GO estimator in the standardized form when the matrix $\frac{1}{\sqrt{n}}\mathbf{X}$ is orthogonal. In this case, we have a closed-form solution for the coefficients as

$$\hat{\beta}_j = \frac{S\left((1 + \lambda_2 d) \hat{\beta}_j^{ls}, \lambda_1\right)}{1 + \lambda_2} = \text{sign}\left((1 + \lambda_2 d) \hat{\beta}_j^{ls}\right) \frac{\left((1 + \lambda_2 d) |\hat{\beta}_j^{ls}| - \lambda_1\right)_+}{1 + \lambda_2}$$

where

$$\text{sign}(z) = \begin{cases} -1, & z < 0 \\ 0, & z = 0 \\ 1, & z > 0 \end{cases}$$

is the signum function and

$$(z)_+ = \begin{cases} z, & z > 0 \\ 0, & z \leq 0 \end{cases}$$

is the positive part function.

3.3. Grouping property

Zou and Hastie (2005) have shown that the elastic net parameter estimates have the grouping effect. A different type of grouping effect can be defined for the GO estimator. To clarify this property for the GO estimator, we first give Theorem 1.

Theorem 1. *Given the response vector \mathbf{y} is centered to have mean zero and the columns of design matrix \mathbf{X} is centered to have mean zero and standard deviation 1. Let $0 \leq d < 1$, $\lambda_1 > 0$ and $\lambda_2 > 0$ parameters are fixed. We assume $\hat{\beta}_i \hat{\beta}_j > 0$. Define the distance*

$$U(i, j) = \frac{1}{\sqrt{\|\mathbf{y}\|_1^2 + n\lambda_2 d^2 \|\hat{\beta}^{ls}\|_2^2}} \left| \left(\hat{\beta}_i - d \hat{\beta}_i^{ls} \right) - \left(\hat{\beta}_j - d \hat{\beta}_j^{ls} \right) \right|, \quad (9)$$

then the inequality

$$U(i, j) \leq \frac{1}{n\lambda_2} \sqrt{2(1-\rho)}$$

holds true where ρ is the sample correlation coefficient between vectors \mathbf{x}_i and \mathbf{x}_j .

Proof. We start the proof with the function given by Equation (4). We first take the derivatives of the function with respect to β_i and β_j and set the results to zero:

$$\left. \frac{\partial Q}{\partial \beta_i} \right|_{\beta_i = \hat{\beta}_i} = -\frac{1}{n} \mathbf{x}_i^\top (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + \lambda_1 \hat{s}_i + \lambda_2 (\hat{\beta}_i - d\hat{\beta}_i^{ls}) = 0 \quad (10)$$

$$\left. \frac{\partial Q}{\partial \beta_j} \right|_{\beta_j = \hat{\beta}_j} = -\frac{1}{n} \mathbf{x}_j^\top (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + \lambda_1 \hat{s}_j + \lambda_2 (\hat{\beta}_j - d\hat{\beta}_j^{ls}) = 0, \quad (11)$$

where \hat{s}_i and \hat{s}_j have the same sign. Then we subtract Equation (10) from Equation (11) and we get

$$\hat{\beta}_j - \hat{\beta}_i - d(\hat{\beta}_j^{ls} - \hat{\beta}_i^{ls}) = \frac{1}{n\lambda_2} (\mathbf{x}_j^\top - \mathbf{x}_i^\top) \hat{\mathbf{r}} \quad (12)$$

where $\hat{\mathbf{r}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. By Cauchy Schwarz inequality for Equation (12), we can write

$$\left| \hat{\beta}_j - \hat{\beta}_i - d(\hat{\beta}_j^{ls} - \hat{\beta}_i^{ls}) \right| \leq \frac{1}{n\lambda_2} \sqrt{\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \|\hat{\mathbf{r}}\|_2^2}. \quad (13)$$

We know that the equality $\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 = 2n(1-\rho)$ holds true because \mathbf{x}_i and \mathbf{x}_j are standardized. Applying this result to Inequality (13), we obtain

$$\left| \hat{\beta}_j - \hat{\beta}_i - d(\hat{\beta}_j^{ls} - \hat{\beta}_i^{ls}) \right| \leq \frac{1}{\sqrt{n}\lambda_2} \sqrt{2(1-\rho)} \|\hat{\mathbf{r}}\|_2. \quad (14)$$

We also know the inequality $Q(\hat{\boldsymbol{\beta}}; d, \lambda_1, \lambda_2) \leq Q(\mathbf{0}; d, \lambda_1, \lambda_2)$ is valid because $\hat{\boldsymbol{\beta}}$ minimizes the function $Q(\hat{\boldsymbol{\beta}}; d, \lambda_1, \lambda_2)$. Therefore, we get

$$\frac{1}{2n} \|\hat{\mathbf{r}}\|_2^2 + \lambda_1 \|\hat{\boldsymbol{\beta}}\|_1 + \frac{\lambda_2}{2} \|\hat{\boldsymbol{\beta}} - d\hat{\boldsymbol{\beta}}^{ls}\|_2^2 \leq \frac{1}{2n} \|\mathbf{y}\|_2^2 + \frac{\lambda_2 d^2}{2} \|\hat{\boldsymbol{\beta}}^{ls}\|_2^2.$$

Considering that the norms are non-negative, we write

$$\|\hat{\mathbf{r}}\|_2^2 \leq \|\mathbf{y}\|_2^2 + n\lambda_2 d^2 \|\hat{\boldsymbol{\beta}}^{ls}\|_2^2. \quad (15)$$

If we take Equations (15) and (14) into account together, we obtain

$$\begin{aligned} \left| \hat{\beta}_j - \hat{\beta}_i - d(\hat{\beta}_j^{ls} - \hat{\beta}_i^{ls}) \right| &\leq \frac{1}{\sqrt{n}\lambda_2} \sqrt{2(1-\rho) (\|\mathbf{y}\|_2^2 + n\lambda_2 d^2 \|\hat{\boldsymbol{\beta}}^{ls}\|_2^2)} \\ &\leq \frac{1}{\sqrt{n}\lambda_2} \sqrt{2(1-\rho)} \sqrt{\|\mathbf{y}\|_1^2 + n\lambda_2 d^2 \|\hat{\boldsymbol{\beta}}^{ls}\|_2^2}. \end{aligned}$$

As a result, the inequality holds and the proof is completed. □

Remark: When two variables, x_i and x_j , are highly negatively correlated, one can consider x_j and $-x_j$ in Theorem 3.1 (Zou and Hastie 2005; Zhou 2013).

The unit-less quantity $U(i, j)$ in Equation (9) allows us to reach some conclusions regarding the differences between the coefficients of explanatory variables i and j for the GO estimator with the shrunken estimates:

- The differences $\hat{\beta}_j - d\hat{\beta}_j^{ls}$ and $\hat{\beta}_i - d\hat{\beta}_i^{ls}$ are approximately equal. Therefore, the approximation speed of the j th GO estimate to the corresponding shrunk estimate is almost equal to the approximation speed of the i th GO estimate to the corresponding shrunk estimate.
- Taking into account that the differences $\hat{\beta}_j - \hat{\beta}_i$ and $d(\hat{\beta}_j^{ls} - \hat{\beta}_i^{ls})$ are approximately equal; that is, the difference between the j th and i th GO estimates is almost equal to the difference between the corresponding shrunk estimates.
- If x_i and x_j are highly correlated, i.e. $\rho = 1$ (if $\rho = -1$ then consider x_i and $-x_j$) the difference between the coefficient paths of x_i and x_j is almost $d(\hat{\beta}_j^{ls} - \hat{\beta}_i^{ls})$. This property also presents that as $d \rightarrow 0$, the grouping property of the GO estimator approaches that of the elastic net.

3.4. Comparison of the GO, lasso and elastic net methods

A graphical comparison of the mentioned methods for different parameter values is given in Figure 1. As can be seen from Figure 1, the curve of the GO estimator has corner points as the same as the case of the lasso and elastic net. These points are the singularity points (i.e., a point without the first derivative) of the penalization function of the GO estimator. These points indicate that the method has variable selection property. Also, the strictly convex structure of the penalization function of the GO estimator is seen in Figure 1, explicitly. Furthermore, the penalization function of the GO estimator can be symmetric (Figure 1 (a) and (c)) or non-symmetric (Figure 1(b)) depending on the coefficient estimates of the shrunk estimator. Besides, the parameter d brings extra flexibility as displayed in Figure 1(c). Figure 1(c) also depicts the parameter space explained in Sec. 3.

Figure 2 shows the operational characteristics of the ridge, lasso, elastic net and GO estimation methods when $\frac{1}{\sqrt{n}}\mathbf{X}$ is orthonormal. The lines in Figure 2 represent the mentioned estimators as different functions of the OLS solution. The elastic net and GO estimators are two-stage procedures although the ridge and lasso are one-stage procedures as mentioned in Zou and Hastie (2005). It is seen from Figure 2 that the interval in which coefficient estimates are zero for the lasso and elastic net are the same, while it is narrow for the GO estimator depending on the parameter d .

We fit the lasso, elastic net and GO estimators to the mpg data set from Henderson and Velleman (1981) when $\alpha = 0.5$ and $d = 0.2$ or $d = 0.9$ for 100 different λ values. The coefficient

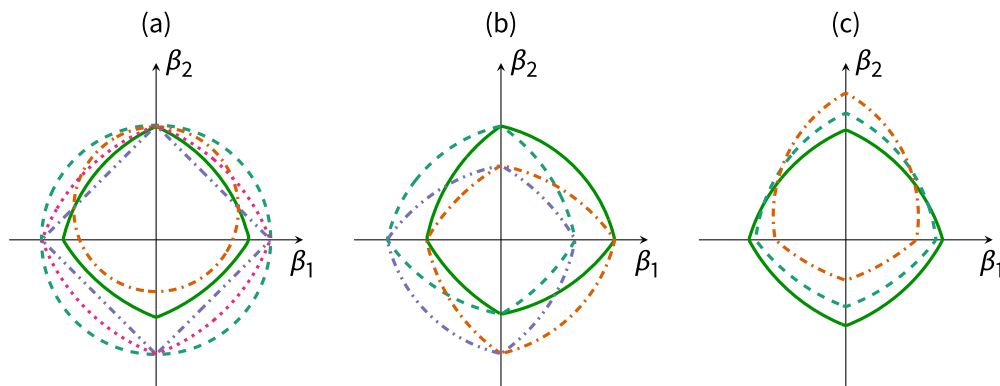


Figure 1. (a) Two dimensional contour plots of ridge penalty (-----), OK penalty (-----), lasso penalty (.....), elastic net penalty (.....) and the GO penalty (——) with $\alpha = 0.5$, $d = 0.9$, $\beta_1^{ls} = 0$, $\beta_2^{ls} = 0.5$; (b) Two dimensional contour plots of the GO penalty for $\beta_1^{ls} = 1$, $\beta_2^{ls} = 1$ (.....), $\beta_1^{ls} = 1$, $\beta_2^{ls} = -1$ (-----), $\beta_1^{ls} = -1$, $\beta_2^{ls} = 1$ (.....), $\beta_1^{ls} = -1$, $\beta_2^{ls} = -1$ (-----) with $\alpha = d = 0.5$. (c) Two dimensional contour plots of the GO penalty for $d = 0.2$ (——), $d = 0.5$ (-----), $d = 0.9$ (.....) with $\alpha = 0.5$, $\beta_1^{ls} = 0$, $\beta_2^{ls} = 1.5$.

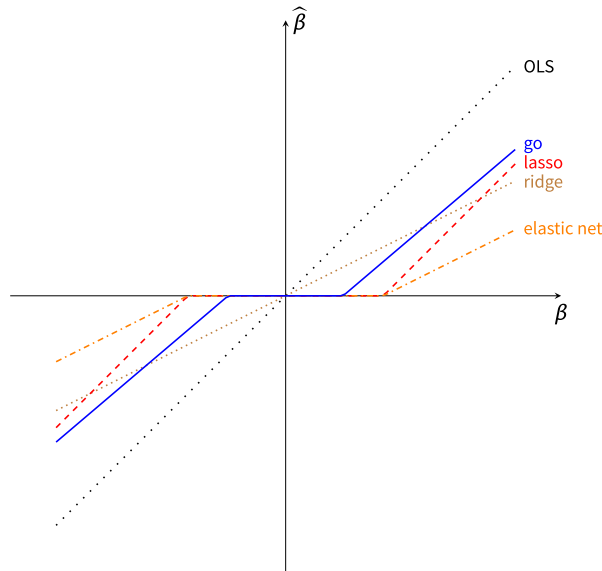


Figure 2. Exact solutions for the ridge (-----), lasso (.....), elastic net (.....) and GO (——) estimators in the orthogonal design where the shrinkage parameters are $\lambda_1 = 2$, $\lambda_2 = 1$ and $d = 0.7$ and (.....) denotes the OLS estimator.

paths of the fitted models are shown in Figure 3. The coefficient paths are plotted as a function of the parameter $\|\beta\|_1 / \max \|\beta\|_1$ for convenience in comparison to different models. Coefficient paths obtained with the lasso show instability while this instability has improved somewhat with the elastic net. When the coefficient paths obtained by the GO estimator are examined, the existence of the knot points continues as in the case of the lasso and elastic net. However, the coefficient paths follow a more stable course. In addition, the points at which the variables enter into the model change depending on the values of d . Moreover, as d increases, coefficient paths become more pronounced for the more significant variables. Thus the GO estimator gives a better result in terms of reducing the instability in the coefficient paths and emphasizing the more significant coefficients.

These conclusions can be seen in Figure 4 where the data were simulated for the sample size of $n = 100$ as follows (Hastie, Tibshirani, and Wainwright 2015):

$$Z_1, Z_2 \sim \mathcal{N}(0, 1) \text{ independent,}$$

$$Y = 3Z_1 - 1.5Z_2 + 2\varepsilon, \text{ with } \varepsilon \sim \mathcal{N}(0, 1),$$

$$X_j = Z_1 + \xi_j/5, \text{ with } \xi_j \sim \mathcal{N}(0, 1), \text{ for } j = 1, 2, 3, \text{ and}$$

$$X_j = Z_2 + \xi_j/5, \text{ with } \xi_j \sim \mathcal{N}(0, 1), \text{ for } j = 4, 5, 6.$$

In this setting, there are two sets of three variables, with a pairwise correlation about 0.97 in each group. We give the coefficient paths of the methods in Figure 4. The horizontal lines in Figure 4(a) and (b) are the OLS estimates while these lines represent the shrunk estimator in Figure 4(c) and (d). All the methods coincide with the OLS solution for $\lambda = 0$ (when $\|\beta\|_1$ has its largest value). According to Figure 4, the coefficients of the lasso do not show good performance since the coefficient paths of the lasso tend to be erratic and it does not capture the relative importance of the variables. Unlike the lasso, the elastic net tends to combine the variables in each group together. Therefore it reflects the relative prominence of the individual variables in the groups. The GO shows a grouping effect in a different manner. As can be seen from Figure 4, the GO estimator drags the coefficient estimates toward the estimates of the shrunk estimator individually. For the small value of d , the pattern is similar to the paths of the elastic net while for the

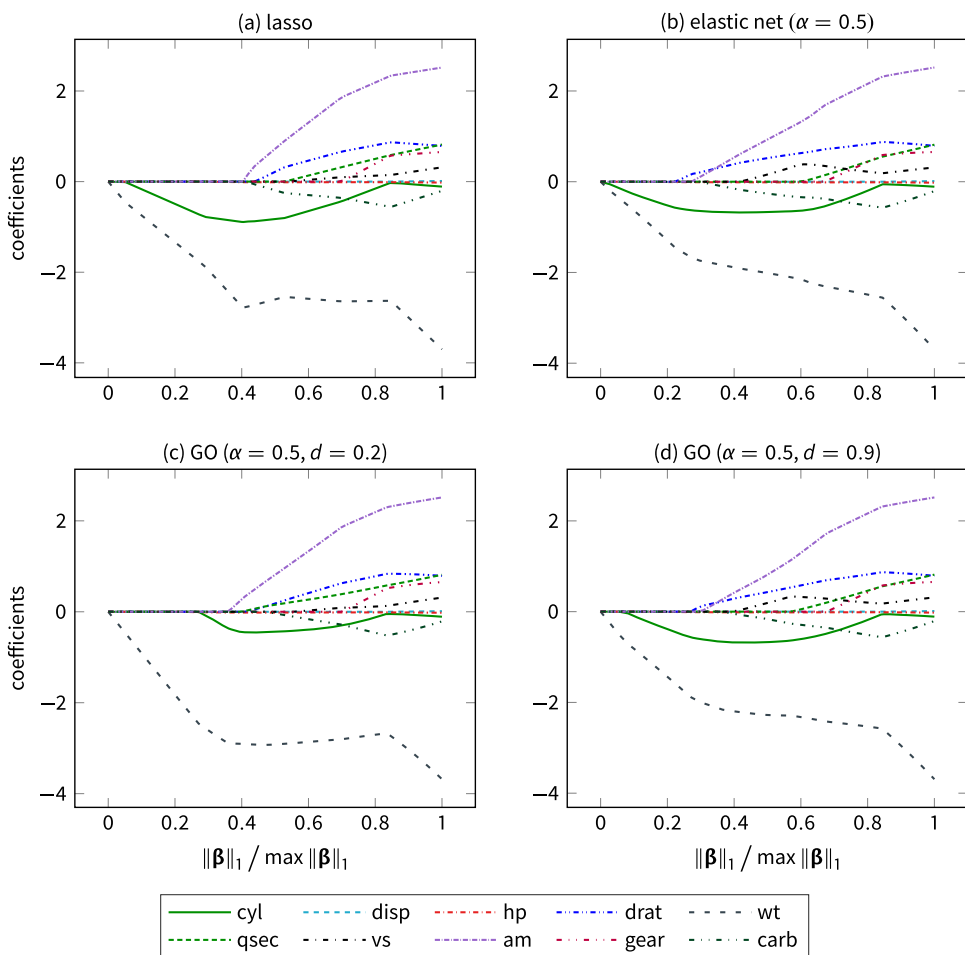


Figure 3. Lasso (a), elastic net (b) and GO (c, d) coefficient estimates as a function of $\|\beta\|_1 / \max \|\beta\|_1$ for the `mpg` data set when $\alpha = 0.5$, $d = 0.2$ and $d = 0.9$ for 100 different values of λ .

large value of d , the individual estimates approach their $d\hat{\beta}^{ls}$ counterparts more quickly while the groups can still be seen easily.

4. Real data analysis

In this section, we investigate the estimators discussed in the paper with the diabetes data (Efron et al. 2004), which is often used in the literature. The data contain 442 observations with 10 explanatory variables. The explanatory variables are age, sex, body mass index (BMI), average blood pressure (BP) and six blood serum measurements (S1 through S6) and the response variable is a quantitative measure of disease progression one year after baseline. The data show that some moderate and high correlations among the explanatory variables, specifically, high correlations are 0.90 between S1, S2 and -0.74 between S3, S4 as shown in Table 1.

The R program is used for the computations. We standardized the explanatory variables to have mean zero and unit variance and the response variable to have mean zero before the analysis. The OLS, ridge, OK, lasso, elastic net and GO estimators were applied to the standardized data. The data are divided into 70% training and 30% test data sets. Tenfold cross-validation is applied to the training data for tuning parameter selection. The standard error estimates of the

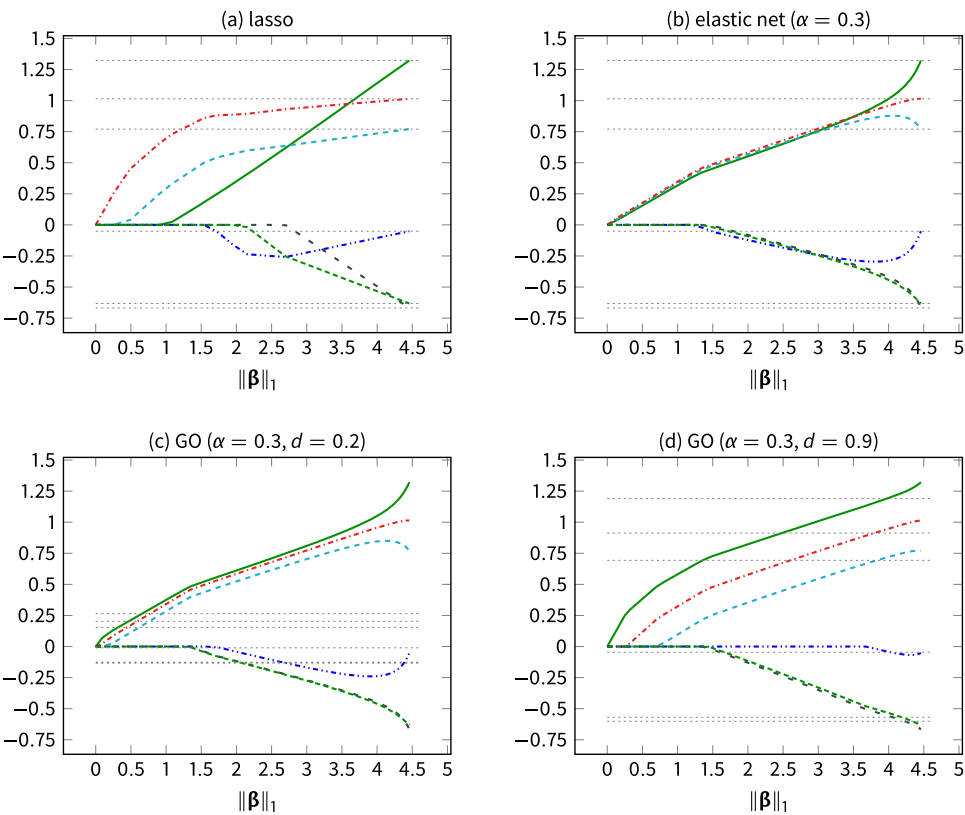


Figure 4. Lasso (a), elastic net (b) and GO (c, d) coefficient estimates as a function of $\|\beta\|_1$ for the simulated data set when $\alpha = 0.3$, $d = 0.2$ and $d = 0.9$ for 100 different values of λ .

Table 1. Correlation table for diabetes data.

	AGE	SEX	BMI	BP	S1	S2	S3	S4	S5	S6	Y
AGE	1	0.17	0.19	0.34	0.26	0.22	−0.08	0.20	0.27	0.30	0.19
SEX		1	0.09	0.24	0.04	0.14	−0.38	0.33	0.15	0.21	0.04
BMI			1	0.40	0.25	0.26	−0.37	0.41	0.45	0.39	0.59
BP				1	0.24	0.19	−0.18	0.26	0.39	0.39	0.44
S1					1	0.90	0.05	0.54	0.52	0.33	0.21
S2						1	−0.20	0.66	0.32	0.29	0.17
S3							1	−0.74	−0.40	−0.27	−0.39
S4								1	0.62	0.42	0.43
S5									1	0.46	0.57
S6										1	0.38
Y											1

model coefficients are obtained by the bootstrap method with 1000 repetitions. To compare the performance of the methods, the prediction mean squared error (PMSE) quantities were calculated on the test set and the selected variables are reported.

The optimal tuning parameters and selected variables by the methods for the diabetes data are shown in Table 2. The estimates of the tuning parameters are obtained by minimizing mean cross-validation error. The lasso selects five variables, SEX, BMI, BP, S3, S5 while the GO estimator selects eight variables including the variables selected by the lasso, S1, S2, S4, additionally, and the elastic net does not make the variable selection. Therefore, the lasso discards positively highly correlated variables S1 and S2 together while selecting one of the negatively

Table 2. Comparison of the methods on diabetes data with parameter values and selected variables.

Method	Parameter values	Variables selected
OLS	—	All the variables
Ridge	$\lambda = 5.745$	All the variables
OK	$\lambda = 4.692, d = 0.24$	All the variables
Lasso	$\lambda = 2.017$	2,3,4,7,9
Elastic net	$\alpha = 0.54, \lambda = 0.131$	All the variables
GO	$\alpha = 0.24, \lambda = 4.762, d = 0.99$	2,3,4,5,6,7,8,9

Table 3. Coefficient estimates, length of the coefficient vectors and PMSE values for each method in diabetes data.

	OLS	Ridge	OK	Lasso	Elastic net	GO
Intercept	150.9159 (3.1626)	150.9159 (3.163)	150.9159 (3.1592)	150.9159 (3.174)	150.9159 (3.1598)	150.9159 (3.1626)
AGE	−0.997 (3.2605)	−0.6364 (2.9521)	−0.7478 (3.0605)	0 (1.6769)	−0.8353 (3.1862)	0 (2.3359)
SEX	−13.1097 (3.4059)	−11.599 (3.0604)	−12.1438 (3.1817)	−9.2736 (3.3944)	−12.9168 (3.3934)	−10.9202 (3.3647)
BMI	22.3371 (3.8076)	21.6054 (3.3324)	21.9366 (3.4971)	22.171 (3.7863)	22.4908 (3.7978)	22.4623 (3.8028)
BP	16.1427 (3.6001)	15.1594 (3.1219)	15.5082 (3.2847)	13.5725 (3.5157)	15.9869 (3.5736)	14.6981 (3.5323)
S1	−37.0154 (24.7961)	−3.5769 (2.924)	−12.1239 (8.1700)	0 (2.4766)	−25.6381 (18.1816)	−17.3065 (15.6423)
S2	25.712 (20.4986)	−1.5712 (3.1489)	5.34 (7.0291)	0 (1.322)	16.3201 (15.2081)	9.349 (13.3582)
S3	3.6807 (12.191)	−9.719 (3.0815)	−6.407 (5.0367)	−12.4039 (3.6598)	−1.1508 (9.2833)	−2.7727 (6.6328)
S4	8.9369 (9.3458)	6.6412 (4.9118)	7.1637 (6.2358)	0 (3.792)	7.8256 (8.6078)	7.0682 (6.2669)
S5	36.4086 (9.4536)	22.4843 (3.7269)	26.2281 (4.5873)	23.5105 (4.2761)	32.2762 (7.5218)	29.4738 (7.0676)
S6	0.1857 (3.5786)	1.6659 (3.1892)	1.1786 (3.3255)	0 (1.9728)	0.286 (3.484)	0 (2.4932)
$\sqrt{\beta^T \beta}$	164.7976	155.7860	156.8519	155.705	160.4139	157.8883
PMSE	3078.3164	3074.4676	3067.7078	3111.6118	3067.4051	3058.2837

highly correlated variables S3 and S4. The GO estimator discards AGE and S6 which are also discarded by the lasso.

The model coefficient estimates and bootstrap standard error estimates of diabetes data are given in Table 3. The problem of incorrect signs in the estimates of S1 and S2 is sorted out by ridge regression. We give norms of the coefficient vectors and PMSE values in the bottom two rows of Table 3. The GO estimator has the smallest PMSE value among all the methods while the lasso has the largest one. This shows that the GO estimator reduces the PMSE for the sake of a small decrease in the variable selection property. The norm of the GO estimator is smaller than the elastic net. In this data set, the GO estimator dominates the elastic net in terms of the variable selection, norm of the coefficient vector and PMSE values. Also, Figure 5 shows the box plots of the bootstrap coefficient estimates of the diabetes data obtained by the estimation methods after 1000 bootstrap iterations. The estimated coefficient of AGE displays symmetric bootstrap distribution while the estimated coefficient of S1, S2, S4 and S6 display skewed bootstrap distributions in view of the lasso method. On the other hand, both the estimated coefficients of AGE and S6 show symmetric bootstrap distributions.

The test MSE values of the estimators versus λ when α and d are fixed at their corresponding optimum values are given in Figure 6. There is not one optimum λ that one estimator outperforms the others. The superiority of one estimator depends on the selected λ value. The GO estimator in this data set dominates the others for λ values larger than 1.

Examination of the data analysis presents that the GO estimator is the best when the tuning parameters are optimized.

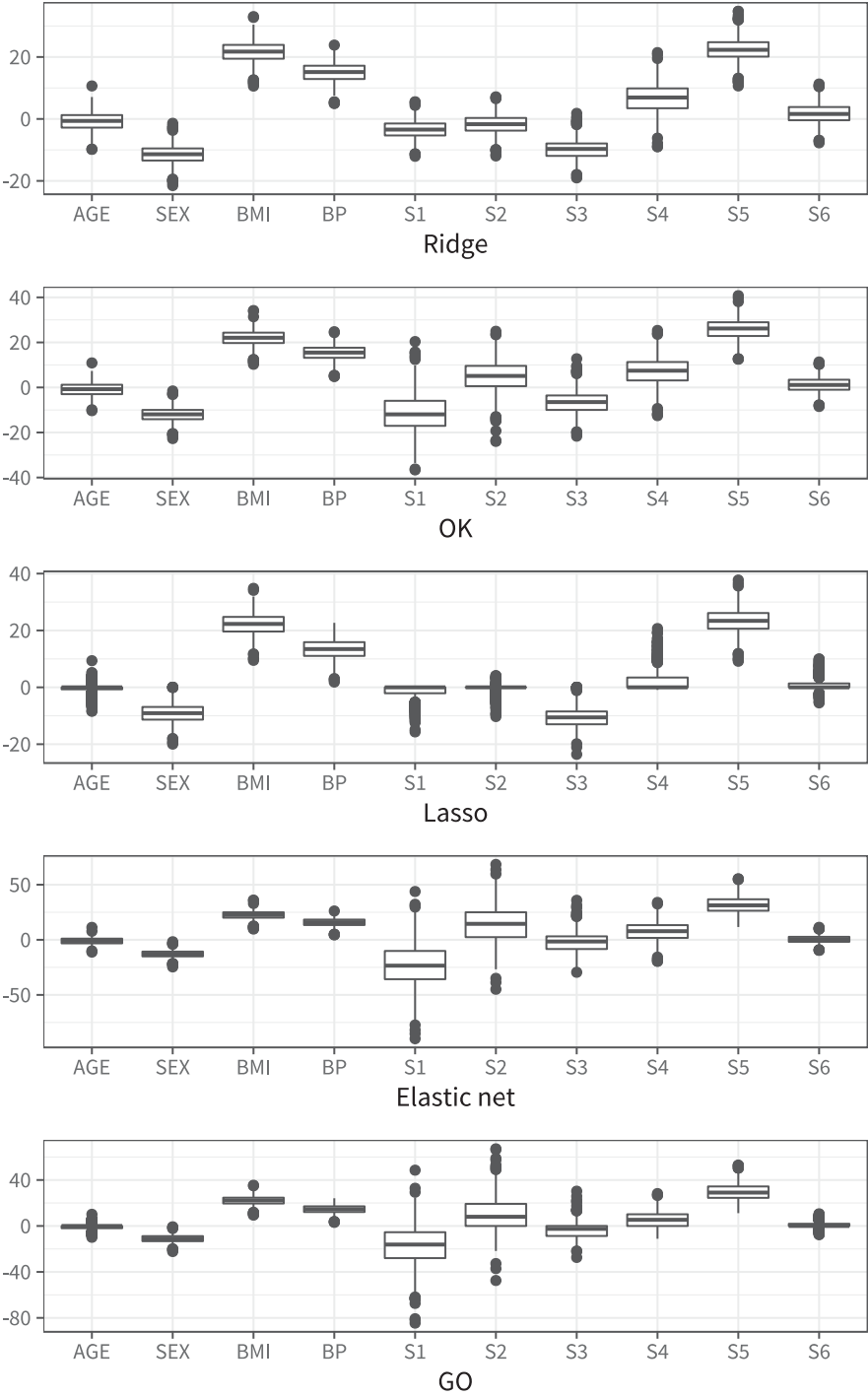


Figure 5. Box plots of bootstrap coefficient estimates of the diabetes data for each of the estimation methods.

5. Simulation studies

In this section, we perform simulation studies to compare the estimators OLS, ridge, lasso, elastic net, OK and GO. We generate data sets consisting of independent training, validation and test

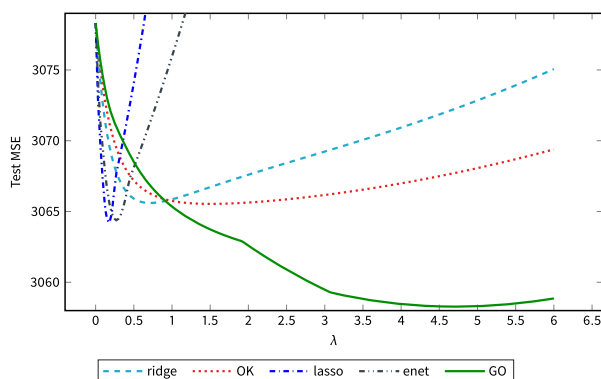


Figure 6. Test MSE values of the diabetes data set for each of the estimation methods.

sets. The training set is used for estimating the coefficients and the validation set is used for estimating the tuning parameters. More clearly, we choose 100 λ values for a grid of $\{d, \alpha\}$ pairs, and fit the models. The model performance is measured by the median of mean squared errors (MSE) where MSE values are computed as

$$(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^\top \mathbf{V}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$$

on the test set, where \mathbf{V} is the covariance matrix of \mathbf{X} and $\hat{\boldsymbol{\beta}}$ is the concerned estimator. (\cdot, \cdot, \cdot) notation indicates the number of observations in the training, validation and test sets, respectively. Moreover, standard error estimates for each estimated test MSE are obtained by the bootstrap method with 1000 repetitions.

The true model for the simulation studies is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}).$$

The signal-to-noise ratio (SNR) for the linear regression is defined as

$$\text{SNR} = \frac{\text{Var}(\mathbf{X}\boldsymbol{\beta})}{\sigma^2}$$

where $\boldsymbol{\beta}$ and σ^2 are estimated by the concerned estimator and low values of SNR imply that the data is sparse, while high values imply that the data is dense (Hastie, Tibshirani, and Friedman 2009). We report the SNR values for all the simulations.

The first two simulation studies are basically from Tibshirani (1996) and the third and fourth simulation studies are from Zou and Hastie (2005). We repeat the simulations for four σ values and three ρ values where ρ is a notation used to evaluate the degree of correlation.

(a) *Simulation Study 1*

In this simulation study, 100 data sets consisting of (20, 20, 200) observations and 8 explanatory variables are generated. The true coefficient vector is $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^\top$. In this setting, we intend to compare the methods when some of the explanatory variables do not relate to the response. We repeat the simulation studies so that the correlation between the explanatory variables \mathbf{x}_i and \mathbf{x}_j is selected to be $\rho_{ij} = \rho^{|i-j|}$ where $\rho = 0.5, 0.7, 0.9, 0.99$ and $\sigma = 3, 5, 15, 25$.

(b) *Simulation Study 2*

In this simulation study, there is no difference in terms of settings with Simulation 1 other than the selected true parameter vector $\boldsymbol{\beta}$. This $\boldsymbol{\beta}$ is selected to be as $\boldsymbol{\beta} = (0.85, 0.85, \dots, 0.85)^\top$. The aim of this setting is to explore if a method has superiority over the other method when all the explanatory variables are related to the response variable.

(c) *Simulation Study 3*

In this simulation study, the setting is the same as Simulation 1 except that the true coefficient vector is

$$\beta = \left(\underbrace{0, \dots, 0}_{10}, \underbrace{2, \dots, 2}_{10}, \underbrace{0, \dots, 0}_{10}, \underbrace{2, \dots, 2}_{10} \right)^\top.$$

where the true coefficient vector has a block-wise structure with 10 repeats in each block.

(d) *Simulation Study 4*

In this simulation study, 100 data sets consisting of (50, 50, 400) observations and 40 explanatory variables are generated. β and σ parameters are set to

$$\beta = \left(\underbrace{3, \dots, 3}_{15}, \underbrace{0, \dots, 0}_{25} \right)^\top$$

and $\sigma = 3, 5, 15, 25$. The explanatory variables \mathbf{x}_i are formed as follows:

$$\mathbf{x}_i = Z_1 + \varepsilon_i^x, Z_1 \sim \mathcal{N}(0, 1), i = 1, 2, \dots, 5,$$

$$\mathbf{x}_i = Z_2 + \varepsilon_i^x, Z_2 \sim \mathcal{N}(0, 1), i = 6, 7, \dots, 10,$$

$$\mathbf{x}_i = Z_3 + \varepsilon_i^x, Z_3 \sim \mathcal{N}(0, 1), i = 11, 12, \dots, 15,$$

$$\mathbf{x}_i \stackrel{iid}{\sim} \mathcal{N}(0, 1), \quad i = 16, 17, \dots, 40 \text{ where } \varepsilon_i^x \stackrel{iid}{\sim} \mathcal{N}(0, 0.01), i = 1, 2, \dots, 15.$$

This data set exhibits the grouping effect.

Tables 4–13 show the simulation results. We report median test MSE values, expected standard error of the median test MSE, reduction rate gained by using the GO estimator and active set size for the estimators in each table. The reduction rate (in percent) is computed as

$$RR = \frac{\text{TMSE}(\hat{\beta}) - \text{TMSE}(\hat{\beta}^{GO})}{\text{TMSE}(\hat{\beta})} \times 100$$

where $\hat{\beta}$ is any estimator to be compared with the GO estimator and $\text{TMSE}(\cdot)$ is the estimated test MSE of the relevant estimator. The aim of using this quantity is to show the relative performance of the GO estimator. We also report active set sizes which represent the number of nonzero coefficient estimates obtained by a method.

We report the results of all the simulation studies as follows:

As a general statement, the GO estimator dominates the other estimators with respect to the test MSE median in all of the simulation studies for all the values of ρ and σ . It competes with the other estimators in the sense of the expected standard error of test MSE median and active set size (for variable selection). We reported the performance results with respect to σ and ρ , separately for Simulation 1, Simulation 2 and Simulation 3. The comments of Simulation 4 are given after the comments of Simulation 1, Simulation 2 and Simulation 3.

5.1. Performance with respect to σ

In Simulation 1, for fixed and small values of σ , the test MSE values of the estimators generally get smaller as the values of ρ increase. However, the reverse is correct for large σ values, yet the GO estimator is less influenced in this case. For very high values of ρ , the gain from the test MSE becomes very large. For large σ values, the reduction rates by the GO estimator in the test MSE median generally increase, as the values of ρ get large. The reductions are very large when both ρ and σ are large. The expected standard errors show a similar pattern in general. The lasso

Table 4. Measurements of the quality for the estimators from Simulation 1 with $\rho = 0.7$ and $\sigma = 3, 5, 15, 25$.

SNR	σ	Measurements	OLS	ridge	OK	lasso	enet	GO
3.0232	3	Median MSE	6.2397	2.8349	2.6518	2.6451	2.2751	1.9305
		Expected SE	0.4469	0.3008	0.1929	0.1637	0.2553	0.2367
		RR	69	32	27	27	15	–
1.0883	5	Active Set Size	8	8	8	5	6	6
		Median MSE	17.3326	5.5443	4.6958	6.3407	4.7104	3.9345
		Expected SE	1.2414	0.9943	0.5308	0.6649	0.687	0.3197
0.1209	15	RR	77	29	16	38	16	–
		Active Set Size	8	8	8	4	6	6
		Median MSE	155.9933	21.4887	14.7010	22.6971	15.6023	12.9109
0.0435	25	Expected SE	11.1726	2.2702	1.6676	1.9088	2.2376	1.4152
		RR	92	40	12	43	17	–
		Active Set Size	8	8	8	2	5	5
		Median MSE	433.3148	25.3838	22.4764	26.3236	23.9441	15.7888
		Expected SE	31.0350	0.6502	1.4724	0.4781	0.9037	1.4407
		RR	96	38	30	40	34	–
		Active Set Size	8	8	8	1	7	4.5

Table 5. Measurements of the quality for the estimators from Simulation 1 with $\rho = 0.9$ and $\sigma = 3, 5, 15, 25$.

SNR	σ	Measurements	OLS	ridge	OK	lasso	enet	GO
4.0327	3	Median MSE	6.2397	1.7804	1.7218	2.0418	1.3544	1.2371
		Expected SE	0.4469	0.1055	0.1330	0.1556	0.1206	0.1089
		RR	80	31	28	39	9	–
1.4518	5	Active Set Size	8	8	8	4	7	6
		Median MSE	17.3326	3.3666	2.9936	3.9421	2.6771	2.3439
		Expected SE	1.2414	0.3862	0.2288	0.4057	0.2873	0.1832
0.1613	15	RR	86	30	22	41	12	–
		Active Set Size	8	8	8	3	6	5.5
		Median MSE	155.9933	16.4650	9.8937	17.7329	12.7782	7.2925
0.0581	25	Expected SE	11.1726	1.9642	1.4736	2.1569	2.7218	0.8456
		RR	95	56	26	59	43	–
		Active Set Size	8	8	8	2	6	4
		Median MSE	433.3148	32.1639	20.2757	32.6794	26.1888	13.2347
		Expected SE	31.0350	2.4084	2.4768	2.353	4.9652	2.628
		RR	97	59	35	60	49	–
		Active Set Size	8	8	8	1	4	4

Table 6. Measurements of the quality for the estimators from Simulation 1 with $\rho = 0.99$ and $\sigma = 3, 5, 15, 25$.

SNR	σ	Measurements	OLS	ridge	OK	lasso	enet	GO
4.6379	3	Median MSE	6.2397	0.8113	0.6814	1.0152	0.6976	0.5444
		Expected SE	0.4469	0.092	0.0705	0.1044	0.0803	0.0563
		RR	91	33	20	46	22	–
1.6696	5	Active Set Size	8	8	8	3	8	5
		Median MSE	17.3326	1.6475	1.1933	2.1464	1.3026	0.8530
		Expected SE	1.2414	0.2011	0.2096	0.1813	0.1975	0.1054
0.1855	15	RR	95	48	29	60	35	–
		Active Set Size	8	8	8	2	7	4
		Median MSE	155.9933	10.8158	6.2407	12.1466	7.9841	3.1547
0.0668	25	Expected SE	11.1726	2.2583	1.3769	2.1924	1.7404	0.903
		RR	98	71	49	74	60	–
		Active Set Size	8	8	8	1	6	3
		Median MSE	433.3148	28.8275	13.9926	33.2426	22.5054	5.0638
		Expected SE	31.035	4.7474	3.9549	4.4507	4.7887	2.4369
		RR	99	82	64	85	77	–
		Active Set Size	8	8	8	1	5.5	3

is the most parsimonious method in point of the variable selection in all cases. The GO estimator usually selects fewer variables than the elastic net for large values of ρ , this is the case in favor of the elastic net for the small values of ρ .

Table 7. Measurements of the quality for the estimators from Simulation 2 with $\rho = 0.7$ and $\sigma = 3, 5, 15, 25$.

SNR	σ	Measurements	OLS	ridge	OK	lasso	enet	GO
2.4761	3	Median MSE	6.2397	1.4869	1.4216	3.2160	1.4162	1.3323
		Expected SE	0.4469	0.1691	0.159	0.1946	0.1705	0.1591
		RR	79	10	6	59	6	–
0.8914	5	Active Set Size	8	8	8	6	8	8
		Median MSE	17.3326	3.4158	3.2862	5.6981	3.1250	2.8514
		Expected SE	1.2414	0.3796	0.39	0.668	0.3313	0.3524
0.0990	15	RR	84	17	13	50	9	–
		Active Set Size	8	8	8	4	8	8
		Median MSE	155.9933	17.9938	11.1979	19.2029	13.6782	8.4004
0.03566	25	Expected SE	11.1726	1.8989	1.749	1.4821	1.8439	0.9231
		RR	95	53	25	56	39	–
		Active Set Size	8	8	8	2	7	6
		Median MSE	433.3148	21.6056	17.2387	21.8910	19.2620	11.0754
		Expected SE	31.035	0.5638	1.9535	0.5583	1.2376	1.3216
		RR	97	49	36	49	43	–
		Active Set Size	8	8	8	1	7	4

Table 8. Measurements of the quality for the estimators from Simulation 2 with $\rho = 0.9$ and $\sigma = 3, 5, 15, 25$.

SNR	σ	Measurements	OLS	ridge	OK	lasso	enet	GO
3.9796	3	Median MSE	6.2397	1.0380	0.9502	2.0092	0.9658	0.8632
		Expected SE	0.4469	0.0957	0.0944	0.2156	0.1048	0.1155
		RR	86	17	9	57	11	–
1.4326	5	Active Set Size	8	8	8	5	8	8
		Median MSE	17.3326	2.3148	2.1148	3.9345	2.2952	1.8010
		Expected SE	1.2414	0.2488	0.2472	0.5321	0.2028	0.2709
0.1592	15	RR	90	22	15	54	22	–
		Active Set Size	8	8	8	4	8	8
		Median MSE	155.9933	13.6949	8.1950	17.6718	11.2320	6.0673
0.0573	25	Expected SE	11.1726	2.4971	1.435	1.9424	1.6909	0.9902
		RR	96	56	26	66	46	–
		Active Set Size	8	8	8	2	7	6
		Median MSE	433.3148	30.7133	17.6941	33.3901	22.8443	10.1068
		Expected SE	31.0350	3.7427	3.3003	2.3025	4.2352	2.5955
		RR	98	67	43	70	56	–
		Active Set Size	8	8	8	1	6	5

Table 9. Measurements of the quality for the estimators from Simulation 2 with $\rho = 0.99$ and $\sigma = 3, 5, 15, 25$.

SNR	σ	Measurements	OLS	ridge	OK	lasso	enet	GO
5.0096	3	Median MSE	6.2397	0.4791	0.3848	1.0229	0.4287	0.3159
		Expected SE	0.4469	0.0748	0.0731	0.1160	0.0697	0.0443
		RR	95	34	18	69	26	–
1.8035	5	Active Set Size	8	8	8	3	8	8
		Median MSE	17.3326	1.2443	0.9243	1.9024	0.9599	0.6123
		Expected SE	1.2414	0.2235	0.1656	0.1913	0.1657	0.077
0.2004	15	RR	96	51	34	68	36	–
		Active Set Size	8	8	8	2	8	5.5
		Median MSE	155.9933	10.1615	5.2711	13.6084	7.7666	1.9951
0.0721	25	Expected SE	11.1726	2.1075	1.4264	2.5371	1.8601	0.9832
		RR	99	80	62	85	74	–
		Active Set Size	8	8	8	1	6	4
		Median MSE	433.3148	27.7260	13.8746	32.0113	22.6510	4.6229
		Expected SE	31.035	5.0445	3.7601	4.5271	4.7876	2.7054
		RR	99	83	67	86	80	–
		Active Set Size	8	8	8	1	5	3

In Simulation 2, where no true coefficient is zero, elastic net and the GO do not perform variable selection for small σ values as it is expected. They exhibit similar behavior to the ridge in this case. However, they eliminate some variables for the large values of σ . The lasso always

Table 10. Measurements of the quality for the estimators from Simulation 3 with $\rho = 0.7$ and $\sigma = 3, 5, 15, 25$.

SNR	σ	Measurements	OLS	ridge	OK	lasso	enet	GO
119.6402	3	Median MSE	6.2028	3.7012	3.2876	23.2018	3.6525	3.2216
		Expected SE	0.1323	0.1836	0.1159	0.3754	0.1185	0.1058
		RR	48	13	2	86	12	–
		Active Set Size	40	40	40	30	34	33
43.0705	5	Median MSE	17.2299	9.1279	6.6686	23.0964	6.9029	6.3136
		Expected SE	0.3674	0.3187	0.2348	0.4393	0.2415	0.2790
		RR	63	31	5	73	9	–
		Active Set Size	40	40	40	28	37	37
4.7856	15	Median MSE	155.0692	35.7752	16.1783	24.5682	16.6294	15.7328
		Expected SE	3.3065	1.9140	0.5841	1.1849	0.6782	0.5541
		RR	90	56	3	36	5	–
		Active Set Size	40	40	40	18	39	39
1.7228	25	Median MSE	430.7477	58.0098	23.5483	28.7546	25.3959	23.0754
		Expected SE	9.1846	2.5869	1.4978	1.8154	1.7715	1.5708
		RR	95	60	2	20	9	–
		Active Set Size	40	40	40	13	39	39

Table 11. Measurements of the quality for the estimators from Simulation 3 with $\rho = 0.9$ and $\sigma = 3, 5, 15, 25$.

SNR	σ	Measurements	OLS	ridge	OK	lasso	enet	GO
150.6436	3	Median MSE	6.2028	3.2102	2.1145	17.2759	2.2114	2.1056
		Expected SE	0.1323	0.1732	0.0620	0.3204	0.0529	0.0538
		RR	66	34	1	88	5	–
		Active Set Size	40	40	40	28	37	37
54.2317	5	Median MSE	17.2299	6.5582	3.3389	17.6652	3.4166	3.2511
		Expected SE	0.3674	0.2945	0.0845	0.4685	0.0700	0.0945
		RR	81	50	3	82	5	–
		Active Set Size	40	40	40	24	39	39
6.0257	15	Median MSE	155.0692	22.5191	6.9073	18.8563	7.4784	6.5900
		Expected SE	3.3065	1.1658	0.6869	1.7428	0.7755	0.5839
		RR	96	71	5	65	12	–
		Active Set Size	40	40	40	13	40	39
2.1693	25	Median MSE	430.7477	37.3455	10.3157	21.2435	12.5836	10.0769
		Expected SE	9.1846	2.3344	1.5244	2.3787	1.5991	1.6101
		RR	98	73	2	53	20	–
		Active Set Size	40	40	40	9	39	39

Table 12. Measurements of the quality for the estimators from Simulation 3 with $\rho = 0.99$ and $\sigma = 3, 5, 15, 25$.

SNR	σ	Measurements	OLS	ridge	OK	lasso	enet	GO
164.7094	3	Median MSE	6.2028	1.2903	0.4209	14.2275	0.4793	0.4195
		Expected SE	0.1323	0.0343	0.0105	0.365	0.0162	0.0126
		RR	93	67	1	97	12	–
		Active Set Size	40	40	40	18	40	40
59.2954	5	Median MSE	17.2299	2.1864	0.5459	14.2300	0.6677	0.5378
		Expected SE	0.3674	0.0893	0.0282	0.4106	0.0508	0.0228
		RR	97	75	1	96	19	–
		Active Set Size	40	40	40	13	40	40
6.5884	15	Median MSE	155.0692	7.5476	1.6531	14.3366	3.1555	1.6089
		Expected SE	3.3065	0.5945	0.255	2.0549	0.3281	0.2311
		RR	99	79	3	89	49	–
		Active Set Size	40	40	40	6	40	39
2.3718	25	Median MSE	430.7477	12.3479	3.3835	16.3130	6.5151	3.0759
		Expected SE	9.1846	0.8440	0.4859	3.3422	0.9173	0.3018
		RR	99	75	9	81	53	–
		Active Set Size	40	40	40	5	40	34.5

carries out the variable selection and, for the large values of σ , the lasso yields very close to the null model. The GO estimator yields more parsimonious models than elastic net in this simulation study for the large values of σ . The GO estimator shows relatively better performance in test

Table 13. Measurements of the quality for the estimators from Simulation 4 with $\sigma = 3, 5, 15, 25$.

SNR	σ	Measurements	OLS	ridge	OK	lasso	enet	GO
72.2968	3	Median MSE	33.7870	7.6657	7.5958	3.5398	2.2854	2.2188
		Expected SE	1.7986	0.4642	0.3776	0.2725	0.236	0.2169
		RR	93	71	71	37	3	–
		Active Set Size	40	40	40	17	22	22
26.0269	5	Median MSE	93.8526	16.8531	16.4937	7.3685	5.9541	5.0700
		Expected SE	4.9899	0.8381	0.7856	0.7131	0.6297	0.5281
		RR	95	70	69	31	15	–
		Active Set Size	40	40	40	15	21	19
2.8919	15	Median MSE	844.6634	64.1817	63.8736	50.8868	45.9231	25.7057
		Expected SE	44.8514	4.1977	3.9976	6.6323	6.1655	2.5761
		RR	97	60	60	49	44	–
		Active Set Size	40	40	40	11	16.5	9
1.0411	25	Median MSE	2346.2817	118.7550	116.3064	130.2332	107.9384	50.9381
		Expected SE	124.556	8.5027	8.1876	15.13	9.5843	5.9959
		RR	98	57	56	61	53	–
		Active Set Size	40	40	40	10	26.5	9

MSE in general, as ρ increases at large σ . In this study, the elastic net dominates the GO estimator with regard to the variable selection. Besides, the reduction rates obtained by the GO estimator seem to be less compared to Simulation 1. Finally, the GO estimator generally produces well estimates of the estimated standard errors of the test MSE median.

In Simulation 3, for fixed σ values, the GO estimator gives the smallest TMSE values, followed by the OK estimator. In this case, GO and OK estimators show more decrease in TMSE than other estimators as the values of the correlation increase. In this study, the active set sizes of the lasso are closer to that of the true coefficient vector than the GO and elastic net estimators. The GO and elastic net estimators yield close active set sizes with the exception of very large σ and ρ values, in which GO yields a more sparse model.

5.2. Performance with respect to ρ

In Simulation 1, the GO estimator dominates the other estimators in test MSE medians for all values of ρ . There is not a general pattern of the reduction rate; however, it is generally large when σ is large for fixed values of ρ . As the values of σ increase, the active set sizes of the lasso are far less than the true active set size, yet the elastic net and the GO yield close to the true active set size. This indicates that the lasso over-shrinks the coefficients for large values of σ ; however, the elastic net and the GO are more robust estimators in this case and yield coefficient estimates with active set size very close to the true size.

In Simulation 2, the estimated test MSE medians and standard error estimates are smaller than the values obtained in Simulation 1 depending on the new true coefficient vector. For large ρ values, the performance of the lasso gets worse as σ increases. This result is consistent with the claim that the ridge can dominate the lasso when $n > p$ if the data set has multicollinearity as Tibshirani (1996) mentioned. In general, the comparison results of the estimators are very similar to that of Simulation 1. We also give similar comments for the active set size to Simulation 1, except that the estimated active set sizes are large for the elastic net and the GO because of the different specification of the true coefficient vector β .

In Simulation 3, the GO estimator has the smallest TMSE values for all values of ρ . The TMSE and expected standard error values increase as σ increases for fixed values of ρ . In this case, as σ increases, the RR value generally increases with the exception of the lasso for all correlation levels and the elastic net with moderate correlation ($\rho = 0.7$). Also, the sparsity level of the lasso increases as σ increases. For the large values of σ , the lasso produces more sparse models

than the true model. The elastic net and GO estimators include more variables than the lasso in this simulation study.

5.3. Results of simulation 4

In Simulation 4, the GO estimator yields the smallest test MSE medians and the reduction rate is very large compared to the other estimators. The expected standard errors of the test MSE median are the lowest for most of the cases. Furthermore, the elastic net and GO estimators picture the true active set size, especially for small values of σ . However, the active set size of the GO estimator for the large values of σ is close to the sizes obtained by the lasso. This situation can be explained by the bias-variance tradeoff: the GO estimator produces more biased estimates compared to the elastic net as σ gets large to overcome the high variance.

6. Modification of the GO estimator to the high-dimensional data

Although the GO estimator has shown success in many situations, it has a limitation in the high-dimensional data ($p > n$) case. This is because the GO estimator relies on the shrunk estimator and the shrunk estimator does not exist for high-dimensional data. In such cases, we modify the objective function (4) by changing $\|\beta - d\hat{\beta}^{ls}\|_2^2$ into $\|\beta - \hat{\beta}^r\|_2^2$. In this case, we have three statements:

- (i) The resulting estimator validates the grouping theorem with a modification that $d\hat{\beta}^{ls}$ is replaced with $\hat{\beta}^r$ in Equation (9). Putting $\hat{\beta}^r$ in $d\hat{\beta}^{ls}$'s place in the objective function leads to a more efficient grouping property since the ridge estimates are more stable than the shrunk estimates in the strong highly correlated case.
- (ii) In this modified version of the GO estimator, the final estimates are shifted toward the ridge solution which may decrease the unnecessary extra bias compared to the elastic net.
- (iii) One difficulty, in this case, is that $\hat{\beta}^r$ remains another tuning parameter, which is more complex to compute. Therefore, the procedure needs extra time consumption than that of the proposed GO estimator.

These intuitive statements of the modified version of the GO estimator are the subject of further study (Genç and Özkale 2021).

7. Conclusions

In this paper, we propose the GO estimator as a new estimation method in penalized regression. The method performs automatic variable selection in a continuous process like the lasso and elastic net. The GO estimator contains the ridge, lasso, elastic net and some other estimators as special cases depending on the values of the parameters. It can be transformed into a lasso problem under a simple transformation of the data, therefore, enjoys the computational advantages of the lasso. The coordinate descent algorithm can be used for the computations of the parameter estimates, which is a very plausible property. The GO estimator has a grouping property, which provides the difference between the coefficients of two highly correlated variables to be proportional to their OLS counterparts by the parameter d ; in other words, shrunk counterparts. This property also ensures in the same situation that the speed of the approximations of the highly correlated variables to their OLS counterparts are approximately equal for sufficiently large d . The analysis of the diabetes data set shows that the GO estimator outperforms the methods that carry out variable selection while the OK estimator has the minimum prediction mean squared error.

We compare the lasso, elastic net, ridge, OK and GO estimators in the context of four simulation studies for various set-ups. The results show that the GO estimator exhibits better performance in general, with respect to the test MSE median with high prediction accuracy. The superiority of the GO estimator becomes prominent when the correlation between the explanatory variables is very high as seen in Simulation 1, Simulation 2 and Simulation 3. The active set sizes obtained by the estimators in the simulation studies show that the GO estimator competes with the lasso and elastic net with regard to the variable selection. Overall, the GO estimator can be seen as an alternative to the methods of variable selection with high prediction performance according to the simulation studies.

Our study shows the GO estimator has considerable properties and, as a new generalization of the lasso, the GO estimator is a competitive method among the methods in penalized regression. One can use the GO estimator instead of the others when the data set is severely correlated.

After the lasso method was proposed, the inadequacies that emerged in the lasso were examined in subsequent literature studies and new methods are still being proposed. Similarly, modifications of the GO estimator, such as group GO, fused GO, adaptive GO, etc. are also possible for future studies.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Murat Genç  <http://orcid.org/0000-0002-6335-3044>

M. Revan Özkale  <http://orcid.org/0000-0001-7085-7403>

References

- Arashi, M., Y. Asar, and B. Yüzbaş I. 2021. Slasso: A scaled lasso for multicollinear situations. *Journal of Statistical Computation and Simulation* 91 (15):3170–83. doi:[10.1080/00949655.2021.1924174](https://doi.org/10.1080/00949655.2021.1924174).
- Dondelinger, F., and S. Mukherjee, Alzheimer's Disease Neuroimaging Initiative. 2020. The joint lasso: High-dimensional regression for group structured data. *Biostatistics (Oxford, England)* 21 (2):219–35. doi:[10.1093/biostatistics/kxy035](https://doi.org/10.1093/biostatistics/kxy035).
- Donoho, D. L., and J. M. Johnstone. 1994. Ideal spatial adaptation by wavelet shrinkage. *Biometrika* 81 (3):425–55. doi:[10.1093/biomet/81.3.425](https://doi.org/10.1093/biomet/81.3.425).
- Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani. 2004. Least angle regression. *The Annals of Statistics* 32 (2): 407–99. doi:[10.1214/009053604000000067](https://doi.org/10.1214/009053604000000067).
- Friedman, J., T. Hastie, and R. Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33 (1):1–22.
- Genç, M., and M. R. Özkale. 2021. Usage of the go estimator in high dimensional linear models. *Computational Statistics* 36 (1):217–39. doi:[10.1007/s00180-020-01001-2](https://doi.org/10.1007/s00180-020-01001-2).
- Gruber, M. H. 2012. Liu and ridge estimators-a comparison. *Communications in Statistics - Theory and Methods* 41 (20):3739–49. doi:[10.1080/03610926.2011.563018](https://doi.org/10.1080/03610926.2011.563018).
- Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The elements of statistical learning: Data mining, inference and prediction*. 2nd ed. New York, NY: Springer.
- Hastie, T., R. Tibshirani, and M. Wainwright. 2015. *Statistical learning with sparsity: The lasso and generalizations*. Boca Raton, FL: CRC press.
- Henderson, H. V., and P. F. Velleman. 1981. Building multiple regression models interactively. *Biometrics* 37 (2): 391–411. doi:[10.2307/2530428](https://doi.org/10.2307/2530428).
- Hoerl, A. E., and R. W. Kennard. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 12 (1):55–67. doi:[10.1080/00401706.1970.10488634](https://doi.org/10.1080/00401706.1970.10488634).
- Jacobucci, R., K. J. Grimm, and J. J. McArdle. 2016. Regularized structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal* 23 (4):555–66. doi:[10.1080/10705511.2016.1154793](https://doi.org/10.1080/10705511.2016.1154793).
- Kong, L. 2022. Fuzzy linear regression model based on adaptive lasso method. *International Journal of Fuzzy Systems* 24 (1):508–18. doi:[10.1007/s40815-021-01156-0](https://doi.org/10.1007/s40815-021-01156-0).

- Lee, J. H., Z. Shi, and Z. Gao. 2021. On lasso for predictive regression. *Journal of Econometrics* 229 (2):322–49.
- Mayer, L. S., and T. A. Willke. 1973. On biased estimation in linear models. *Technometrics* 15 (3):497–508. doi:[10.1080/00401706.1973.10489076](https://doi.org/10.1080/00401706.1973.10489076).
- Özkale, M. R., and S. Kaç İranlar. 2007. The restricted and unrestricted two-parameter estimators. *Communications in Statistics - Theory and Methods* 36 (15):2707–25. doi:[10.1080/03610920701386877](https://doi.org/10.1080/03610920701386877).
- Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 58:267–88.
- Tibshirani, R., and J. Friedman. 2020. A pliable lasso. *Journal of Computational and Graphical Statistics: a Joint Publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America* 29 (1):215–25. doi:[10.1080/10618600.2019.1648271](https://doi.org/10.1080/10618600.2019.1648271).
- Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight. 2005. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (1):91–108. doi:[10.1111/j.1467-9868.2005.00490.x](https://doi.org/10.1111/j.1467-9868.2005.00490.x).
- Xu, X., X. Li, and J. Zhang. 2020. Regularization methods for high-dimensional sparse control function models. *Journal of Statistical Planning and Inference* 206:111–26. doi:[10.1016/j.jspi.2019.09.007](https://doi.org/10.1016/j.jspi.2019.09.007).
- Yuan, M., and Y. Lin. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68 (1):49–67. doi:[10.1111/j.1467-9868.2005.00532.x](https://doi.org/10.1111/j.1467-9868.2005.00532.x).
- Zhou, D.-X. 2013. On grouping effect of elastic net. *Statistics & Probability Letters* 83 (9):2108–12. doi:[10.1016/j.spl.2013.05.014](https://doi.org/10.1016/j.spl.2013.05.014).
- Zou, H. 2006. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101 (476):1418–29. doi:[10.1198/016214506000000735](https://doi.org/10.1198/016214506000000735).
- Zou, H., and T. Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2):301–20. doi:[10.1111/j.1467-9868.2005.00503.x](https://doi.org/10.1111/j.1467-9868.2005.00503.x).