machine learning & knowledge extraction

MDPI

Review

High-Dimensional LASSO-Based Computational Regression Models: Regularization, Shrinkage, and Selection

Frank Emmert-Streib [1,2,*] and Matthias Dehmer [3,4,5]

1 Predictive Society and Data Analytics Lab, Faculty of Information Technology and Communication Sciences, Tampere University, 33720 Tampere, Finland
2 Institute of Biosciences and Medical Technology, 33520 Tampere, Finland
3 Institute for Intelligent Production, Faculty for Management, University of Applied Sciences Upper Austria, 4400 Steyr Campus, Austria; matthias.dehmer@fh-steyr.at
4 Department of Mechatronics and Biomedical Computer Science, UMIT, 6060 Hall in Tyrol, Austria
5 College of Computer and Control Engineering, Nankai University, Tianjin 300071, China
* Correspondence: v@bio-complexity.com; Tel.: +358-50-301-5353

Received: 9 December 2018; Accepted: 11 January 2019; Published: 14 January 2019

Abstract: Regression models are a form of supervised learning methods that are important for machine learning, statistics, and general data science. Despite the fact that classical ordinary least squares (OLS) regression models have been known for a long time, in recent years there are many new developments that extend this model significantly. Above all, the *least absolute shrinkage and selection operator* (LASSO) model gained considerable interest. In this paper, we review general regression models with a focus on the LASSO and extensions thereof, including the adaptive LASSO, elastic net, and group LASSO. We discuss the regularization terms responsible for inducing coefficient shrinkage and variable selection leading to improved performance metrics of these regression models. This makes these modern, computational regression models valuable tools for analyzing high-dimensional problems.

Keywords: machine learning; statistics; regression models; LASSO; regularization; high-dimensional data; data science; shrinkage; feature selection

1. Introduction

The increasing digitalization of our society and progress in the development of new measurement devices has led to a flood of data. For instance, data from social media enable the development of methods for their analysis to address relevant questions in the computational social sciences [1,2]. In biology or the biomedical sciences, novel sequencing technologies enable the generation of high-throughput data from all molecular levels, including mRNAs, proteins, and DNA sequences [3,4]. Depending on the characteristics of the data, appropriate analysis methods need to be selected for their interrogations. One of the most widely used analysis methods is regression models [5,6]. Put simply, this type of method performs a mapping from a set of input variables to output variables. In contrast to classification methods, the output variables for regression models assume real values. Due to the fact that many application problems come in this form, regression models find widespread applications across many fields, e.g., [7–11].

In recent years, several new regression models have been introduced that extend classical regression models significantly. The purpose of this paper is to review such regression models, with a special

focus on the *least absolute shrinkage and selection operator* (LASSO) model [12] and extensions thereof. Specifically, in addition to the LASSO, we will discuss the non-negative garrotte [13], Dantzig selector [14], Bridge regression [15], adaptive LASSO [16], elastic net [17], and group LASSO [18].

Interestingly, despite the popularity of the LASSO, there are only very few reviews available about this model. In contrast to previous reviews about this topic [9,19–21], our focus is different with respect to the following points. First, we focus on the LASSO and advanced models related to the LASSO. Our aim is not to cover all regression models but regularized regression models centered around the LASSO (the concept of *regularization* has been introduced by Tikhonov to approximate ill-posed inverse problems [22,23]). Second, we present the necessary technical details of the methods to the level where they are needed for a deeper understanding. However, we do not present all details especially if they are related to the proof of properties. Third, our explanations aim at an intermediate level of the reader by providing also background information frequently omitted in advanced texts. This should ensure that our review is useful for a broad readership from many areas. Fourth, we use a data set from economics to discuss properties of the methods and to cross-discuss differences among them. Fifth, we will provide information about the practical application of the methods by providing information about availability of implementations for the statistical programming language R. In general, there are many software packages available in different implementations and programming languages but we focus on R because the more statistics oriented literature favors this programming language.

This paper is organized as follows. In the next section, we present general preprocessing steps we use before a regression analysis and we discuss an example data set we use to demonstrate the different models. Thereafter, we discuss ordinary least squares regression and ridge regression because we assume that not all readers will be familiar with these models but an understanding of these is necessary in order to understand more advanced regression models. Then we discuss the non-negative garrotte, LASSO, Bridge regression, Dantzig selector, adaptive LASSO, elastic net, and group LASSO, with a special focus on the regularization term. The paper finishes with a brief summary of the methods and conclusions.

2. Preprocessing of Data and Example Data

We begin our paper by briefly providing some statistical preliminaries needed for the regression models. First, we discuss some preprocessing steps used for standardizing the data for all regression models. Second, we discuss data we are using to demonstrate the differences of the different regression models.

2.1. Preprocessing

Let us assume we have data of the form $(\boldsymbol{x}_i, y_i)$ with $i \in \{1, \ldots, n\}$, where n is the number of samples. The vector $\boldsymbol{x}_i$ corresponds to the predictor variables for sample i, whereas $\boldsymbol{x}_i = (X_{i1}, \ldots, X_{ip})^T$ and p is the number of predictors, furthermore y_i is the response variable. We denote by $\boldsymbol{y} \in \mathbb{R}^n$ the vector of response variables and by $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ the predictor matrix. The vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ gives the regression coefficients.

The predictors and response variable shall be standardized, which means:

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} X_{ij} = 0 \quad \text{for all } j \tag{1}$$

$$\bar{s}_j^2 = \frac{1}{n} \sum_{i=1}^{n} X_{ij}^2 = 1 \quad \text{for all } j \tag{2}$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = 0 \tag{3}$$

Here $\bar{x}_j$ and $\bar{s}_j^2$ are the mean and variance of the predictor variables and $\bar{y}$ is the mean of the response variables.

In order to study the regularization of regression models, we need to solve optimization problems which are formulated in terms of norms. For this reason, we review in the following the norms needed for the subsequent sections. For a real vector $\boldsymbol{x} \in \mathbb{R}^n$ and $q \geq 1$ the Lq-norm is defined by

$$\|\boldsymbol{x}\|_q = \Big(\sum_{i=1}^{n} |x_i|^q \Big)^{\frac{1}{q}}. \tag{4}$$

For the special case $q = 2$ one obtains the L2-norm (also known as Euclidean norm) and for $q = 1$ the L1-norm. Interestingly, for $q < 1$ Equation (4) is defined but no longer a norm in the mathematical sense.

We will revisit the L2-norm when discussion ridge regression and the L1-norm for the LASSO. The infinity norm, also called maximum norm, is defined by

$$\|\boldsymbol{x}\|_\infty = \max_i |x_i|. \tag{5}$$

This norm is used by the Danzig selector.

For $q = 0$ one obtains the L0-norm which corresponds to the number of non-zero elements, i.e.,

$$\|\boldsymbol{x}\|_0 = \#\text{non-zero elements}. \tag{6}$$

2.2. Data

In order to provide some practical examples for the regression models, we use a data set from [24]. The whole data set consists of 156 samples for 93 economic variables about inflation indexes and macroeconomic variables of the Brazilian econnomy. From these we select 7 variables to predict the Brazilian inflation. We focus on 7 variables because these are sufficient to demonstrate the regularization, shrinkage, and selection of the different regression models we discuss in the following sections. Using more variables leads quickly to cumbersome models that require much more effort for their understanding without providing more insights regarding the focus of our paper.

3. Ordinary Least Squares Regression

We begin our discussion by formulating a multiple regression problem,

$$y_i = \sum_{j=1}^{p} X_{ij}\beta_j + \epsilon_i. \tag{7}$$

Here X_{ij} are p predictor variables that are linearly mapped onto the response variable y_i for sample i. The mapping is defined by the p regression coefficients β_j. Furthermore, the mapping is effected by a noise term ϵ_i assuming values in $\sim N(0, \sigma^2)$. The noise term summarizes all kinds of uncertainties, e.g., measurement errors.

In order to see the similarity between a multiple linear regression, having p predictor variables, and a simple linear regression, having one predictor variable, one can write Equation (7) in the form:

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_i \tag{8}$$

Here $\boldsymbol{x}_i^T\boldsymbol{\beta}$ is the inner product (scalar product) between the two p-dimensional vectors $\boldsymbol{x}_i = (X_{i1}, \ldots, X_{ip})^T$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$. One can further summarize Equation (8) for all samples $i \in \{1, \ldots, n\}$ by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{9}$$

Here the noise terms assumes the form $\boldsymbol{\epsilon} \sim N(0, \sigma^2\boldsymbol{I}_n)$ whereas $\boldsymbol{I}_n$ is the $\mathbb{R}^{n\times n}$ identity matrix.
The solution of Equation (9) can be formulated as an optimization problem given by

$$\hat{\boldsymbol{\beta}}^{OLS} = arg\min \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2. \tag{10}$$

The ordinary least squares (OLS) solution of Equation (10) can be analytically calculated assuming X has full column rank, which implies that X^TX is positive definite, and is given by

$$\hat{\boldsymbol{\beta}}^{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}. \tag{11}$$

If X has not full column rank the solution cannot be uniquely determined.

Limitations

Least squares regression can perform very badly when there are outliers in the data. For this reason it can be very helpful in performing outlier detection in the data before the analysis is performed and removing the outliers from the data before applying it to the regression model. A reason why least squares regression is so sensitive to outliers is that this model does not perform any form of coefficient shrinkage of regression coefficients as, e.g., the LASSO. For this reasons coefficients can become very large as a result of such outliers without a limiting mechanism built into the model.

Another factor that can lead to a bad performance is the correlation between predictor variables. The disadvantage of the regression model is that it does not perform any form of variable selection to reduce the numbers of predictor variables as, e.g., ridge regression or LASSO. Instead, it uses the variables specified as input to the model.

The third factor that can reduce the performance is called heteroskedasticity or heteroscedasticity. It refers to varying (non-constant) variances of the error term in dependence on the sampling region. One particular problem caused by heteroskedasticity is that it leads to inefficient and biased estimates of the OLS standard errors and, hence, results in biased statistical tests of the regression coefficients [25].

In summary, ordinary least squares regression neither performs shrinkage nor variable selection, which can lead to problems as discussed above. For this reason, advanced regression models have been introduced to guard against such problems.

4. Ridge Regression

The motivation for improving OLS is the fact that the estimates from such models have often a low bias but a large variance. This is related to the prediction accuracy of a model because it is known that either by shrinking the values of regression coefficients or by setting coefficients to zero the accuracy of a prediction can be improved [26]. The reason for this is that by introducing some estimation bias the variance can be actually reduced.

Ridge regression has been introduced by [27]. The model can be formulated as follows.

$$\hat{\boldsymbol{\beta}}^{RR} = arg\min\left\{\frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \sum_j x_{ij}\beta_j\right)^2 + \lambda\|\boldsymbol{\beta}\|_2^2\right\} \quad (12)$$

$$= arg\min\left\{\frac{1}{2n}\mathrm{RSS}(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_2^2\right\} \quad (13)$$

$$= arg\min\left\{\frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2\right\} \quad (14)$$

Here $\mathrm{RSS}(\boldsymbol{\beta})$ is the residual sum of squares (RSS) called the loss of the model, $\lambda\|\boldsymbol{\beta}\|_2^2$ is the regularization term or penaltyd, and λ is the tuning or regularization parameter. The parameter λ controls the shrinkage of coefficients. The L2-penalty in Equation (12) is sometimes also called Tikhonov regularization.

Ridge regression has an analytical solution which is given by

$$\hat{\boldsymbol{\beta}}^{RR}(\lambda) = \left(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}. \quad (15)$$

Here $\boldsymbol{I}_p$ is the $p \times p$ identity matrix. A problem of OLS is that if $rank(\boldsymbol{X}) < p$, then $\boldsymbol{X}^T\boldsymbol{X}$ does not have an inverse. However, a non-zero regularization parameter λ leads usually to a matrix $\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p$, for which an inverse exists.

In Figure 1, we show numerical examples for the economic data set. Specifically, in Figure 1A–C we show the regression coefficients in dependence on λ because the solution in Equation (15) depends on the tuning parameter. On the top of each figure the numbers of non-zero regression coefficients are shown. One can see that for decreasing value of λ the values of the regression coefficients are decreasing. This is the shrinkage effect of the tuning parameter. Furthermore, one can see that none of the coefficients becomes zero. Instead, all regression coefficients assume small but non-zero values. These observations are characteristics for general results from ridge regression [26].

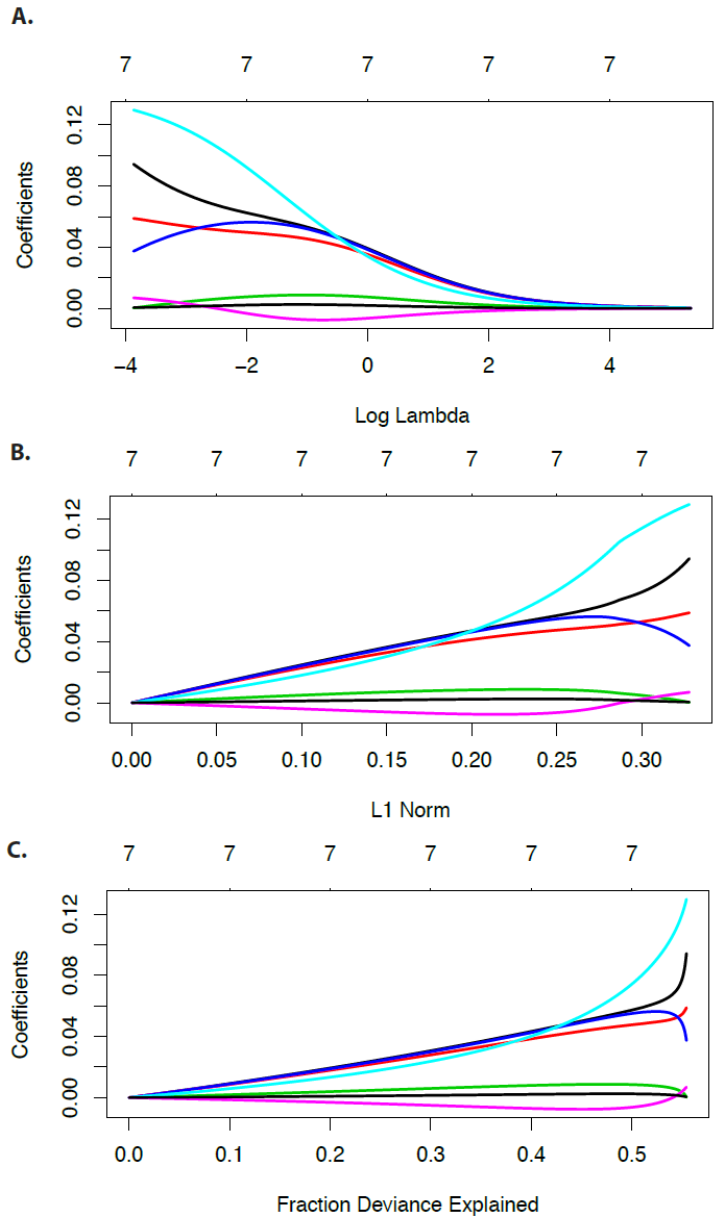

Figure 1. Coefficient paths for the Ridge regression model in dependence on $\log(\lambda)$ (**A**), the L1-norm **(B)**, and the fraction of deviance explained (**C**).

In Figure 2, we show results for the Akaike information criterion (AIC) (Figure 2A), the Bayesian information criterion (BIC) (Figure 2B), and the mean-squared error (Figure 2C). Again, the numbers on top of the figures give the number of non-zero regression coefficients. Each criterion can be used to identify an optimal λ value (see the vertical dashed lines). However, using AIC or BIC would lead to λ values that do not perform a shrinkage of the coefficients (see Figure 1A). In contrast, the mean-squared error suggests a smaller value of the tuning parameter that indeed shrinks the coefficients.

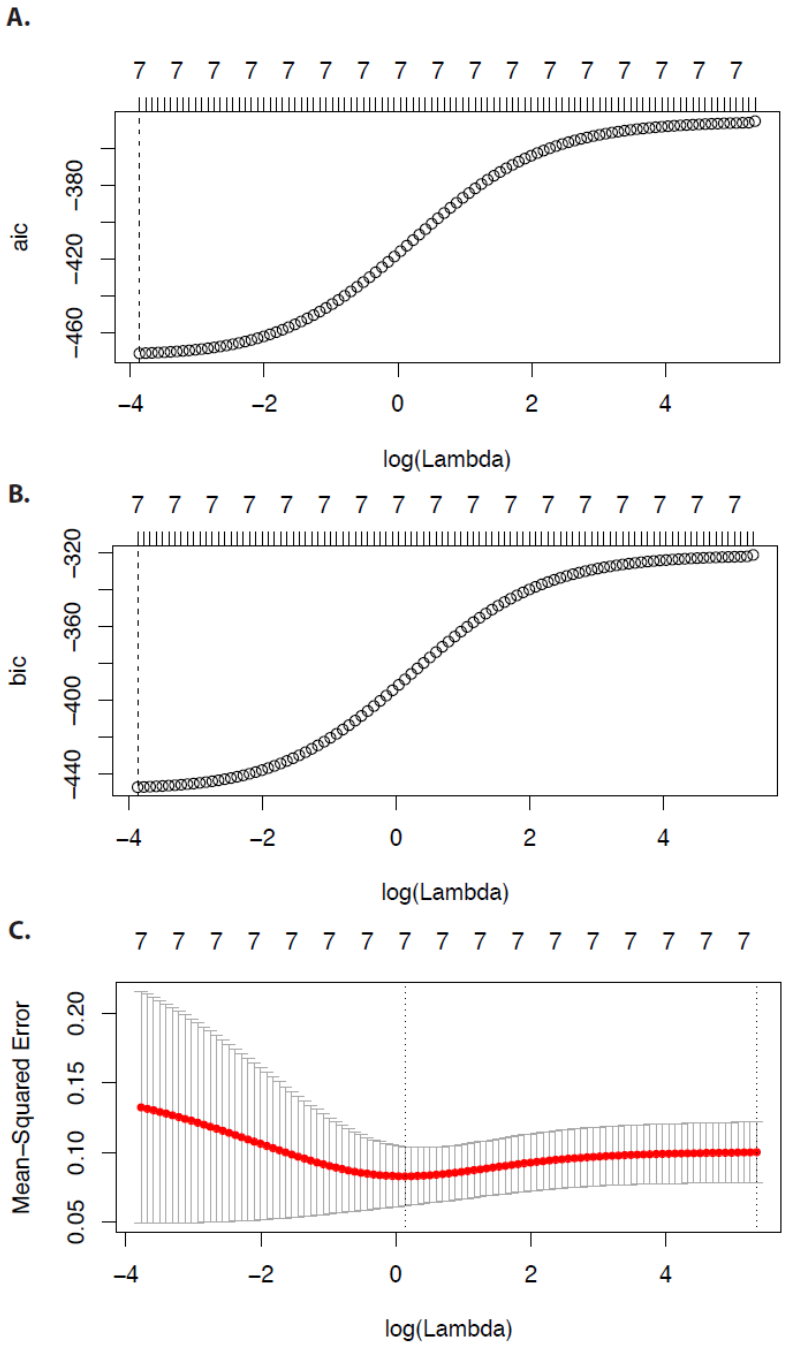

Figure 2. Ridge regression model. (**A**) Akaike information criterion (AIC). (**B**) Bayesian information criterion (BIC). (**C**) Mean-squared error. All results are shown in dependence on the regularization parameter $\log(\lambda)$.

Overall, the advantages of a ridge regression is that it can reduce the variance by paying the price of an increasing bias. This can improve the prediction accuracy of a model. This works best in situations where the OLS estimates have a high variance and for the cases $p < n$.

A disadvantage is that ridge regression does not shrink coefficient to zero and, hence, does not perform variable selection. Another motivating factor for improving upon OLS is that by reducing the number of predictors the interpretation of a model becomes easier because one can focus on the relevant variables of the problem.

R Package

Ridge regression can be performed using the *glmnet* R package [28]. This package is flexible allowing to perform also other types of regularized regression models (see LASSO and adaptive LASSO).

5. Non-Negative Garotte Regression

The next model we discuss, the non-negative garotte, has been mentioned as a motivation for the introduction of the LASSO [12]. For this reason we discuss it before the LASSO. The non-negative garotte has been introduced by [13] and is given by

$$\hat{\beta} \quad = \quad arg\min \left\{ \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{d}\|_2^2 + \lambda \sum_{j=1}^{p} d_j \right\}, \tag{16}$$

for $\boldsymbol{d} = (d_1, \ldots, d_p)^T$ with $d_j > 0$ for all j. The regression is formulated for the scaled variables $\boldsymbol{Z}$ given by $Z_j = X_j \hat{\beta}_j^{OLS}$. That means the model, first, estimates ordinary least squares parameter $\hat{\beta}_j^{OLS}$ for the unregularized regression (Equation (10)) and then performs in a second step a regularized regression for the scaled predictors $\boldsymbol{Z}$.

The non-negative garotte estimate can be expressed with the OLS regression coefficient and the regularization coefficients by [29]

$$\hat{\beta}_j^{NNG}(\lambda) = d_j(\lambda)\hat{\beta}_j^{OLS}. \tag{17}$$

Breiman showed that the non-negative garotte has consistently lower prediction error than subset selection and is competitive with ridge regression except when the true model has many small non-zero coefficients. A disadvantage of the non-negative garotte is its explicit dependency on the OLS estimates [12].

6. LASSO

The LASSO (least absolute shrinkage and selection operator) has been made popular by Robert Tibshirani in 1996 [12], but it had previously appeared in the literature see, e.g., [30,31]. It is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical regression model.

The LASSO estimate of $\hat{\beta}$ is given by:

$$\hat{\boldsymbol{\beta}} = arg\min \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \sum_j \beta_j x_{ij}\right)^2 \right\} \tag{18}$$

$$\text{subject to:} \|\boldsymbol{\beta}\|_1 \leq t \tag{19}$$

Equation (18) is called the constrained form of the regression model. In Equation (19) t is a tuning parameter (also called regularization parameter or penalty parameter) and $\|\boldsymbol{\beta}\|_1$ the L1-norm (see Equation (4)).

It can be shown that Equation (18) can be written in the Lagrange form given by:

$$\hat{\boldsymbol{\beta}} = arg\min\left\{\frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\sum_j \beta_j x_{ij}\right)^2+\lambda\|\boldsymbol{\beta}\|_1\right\} \quad (20)$$

$$= arg\min\left\{\frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda\|\boldsymbol{\beta}\|_1\right\} \quad (21)$$

The relation between both forms holds due to the duality and the KKT (Karush-Kuhn-Tucker) conditions. Furthermore, for every $t > 0$ there exists a $\lambda > 0$ such that both equations lead to the same solution [26].

In general, the LASSO lacks a closed form solution because the objective function is not differentiable. However, it is possible to obtain closed form solutions for the special case of an orthonormal design matrix.

In the LASSO regression model Equation (21), λ is a parameter that needs to be estimated. This is accomplished by cross-validation. Specifically, for each fold F_k the mean-squared error is estimated by

$$e(\lambda)_k = \frac{1}{\#F_k}\sum_{j\in F_k}(y_j-\hat{y}_j)^2. \quad (22)$$

Here $\#F_k$ is the number of samples in set F_k. Then the average over all K folds is taken,

$$CV(\lambda) = \frac{1}{K}\sum_{k=1}^{K} e(\lambda)_k. \quad (23)$$

This is called the cross-validation mean-squared error. For obtaining an optimal λ from $CV(\lambda)$, two approaches are common. The first estimates the λ that minimizes the function $CV(\lambda)$,

$$\hat{\lambda}_{min} = arg\min CV(\lambda). \quad (24)$$

The second approach, first, estimates $\hat{\lambda}_{min}$ and then identifies the maximal λ that has a cross-validation MSE (mean squared error) smaller than $CV(\hat{\lambda}_{min}) + SE(\hat{\lambda}_{min})$, given by

$$\hat{\lambda}_{1se} = \max_{CV(\lambda)\leq CV(\hat{\lambda}_{min})+SE(\hat{\lambda}_{min})} \lambda. \quad (25)$$

6.1. Example

In Figures 3 and 4 we show examples for the economy data. In Figure 3 we show coefficient paths for the LASSO regression model in dependence on $\log(\lambda)$ (A), the L1-norm (B), and the fraction of deviance explained (C).

In Figure 4 we show the Akaike information criterion, the Bayesian information criterion (see [32]), and the Mean-squared error of λ.

6.2. Explanation of Variable Selection

From Figure 3 one can see that decreasing values of λ lead to the shrinkage of the regression coefficients and some of these even become zero. To understand this behavior, we depict in Figure 5A,B a two-dimensional LASSO (A) and ridge regression (B) model. The regularization term of each regression model is depicted in blue, corresponding to the diamond shape for the L1-norm and the circle for the L2-norm. The solution of the optimization problem is given by the intersection of the ellipsis and the

boundary of the penalty shapes. These intersections are highlighted by a green point for the LASSO and a blue point for the ridge regression. In order to shrink a coefficient to zero an intersection needs to occur alongside the two coordinate axis. For the shown situation this is only possible for the LASSO but not for ridge regression. In general, the probability for a LASSO to shrink a coefficient to zero is much larger than for the ridge regression.

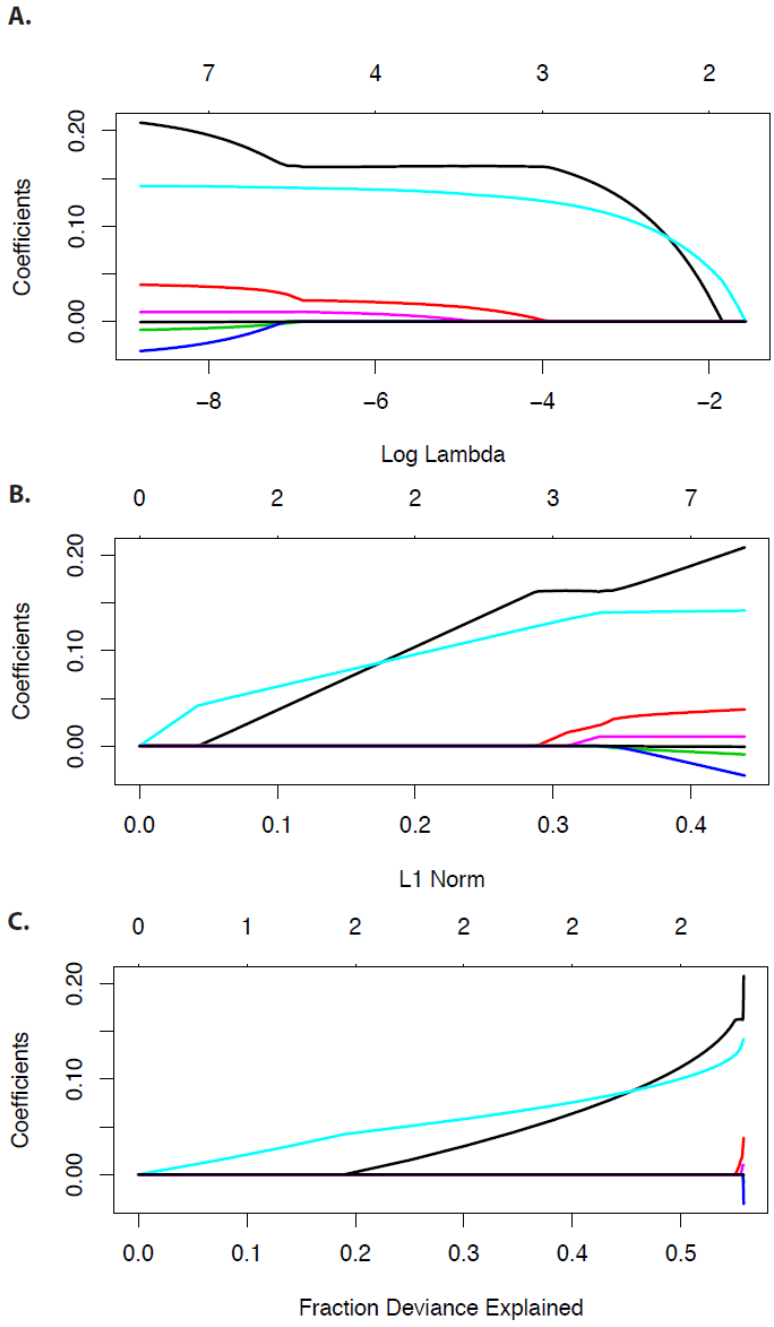

Figure 3. Coefficient paths for the least absolute shrinkage and selection operator (LASSO) regression model in dependence on $\log(\lambda)$ (**A**), the L1-norm (**B**), and the fraction of deviance explained (**C**).

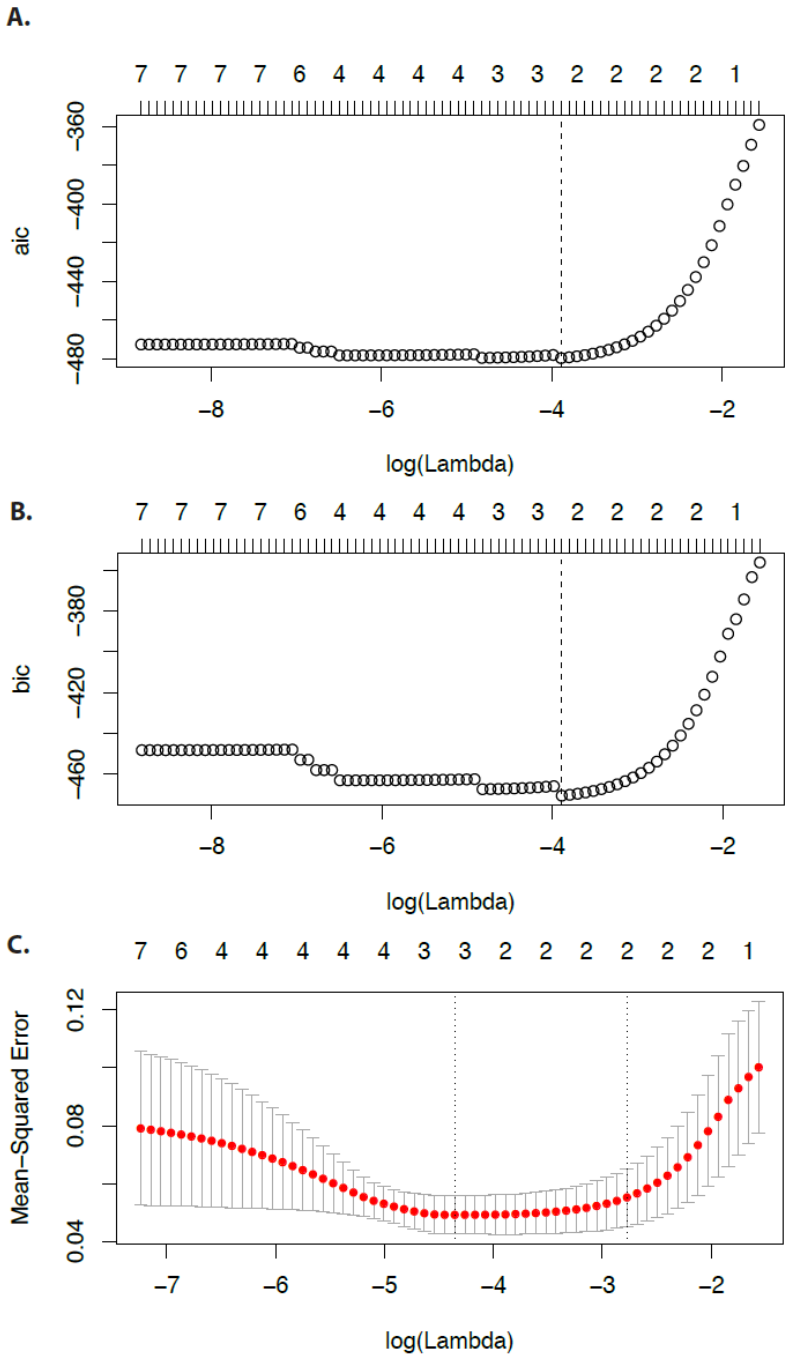

Figure 4. LASSO regression model. (**A**) Akaike information criterion (AIC). (**B**) Bayesian information criterion (BIC). (**C**) Mean-squared error. All results are shown in dependence on the regularization parameter $\log(\lambda)$.

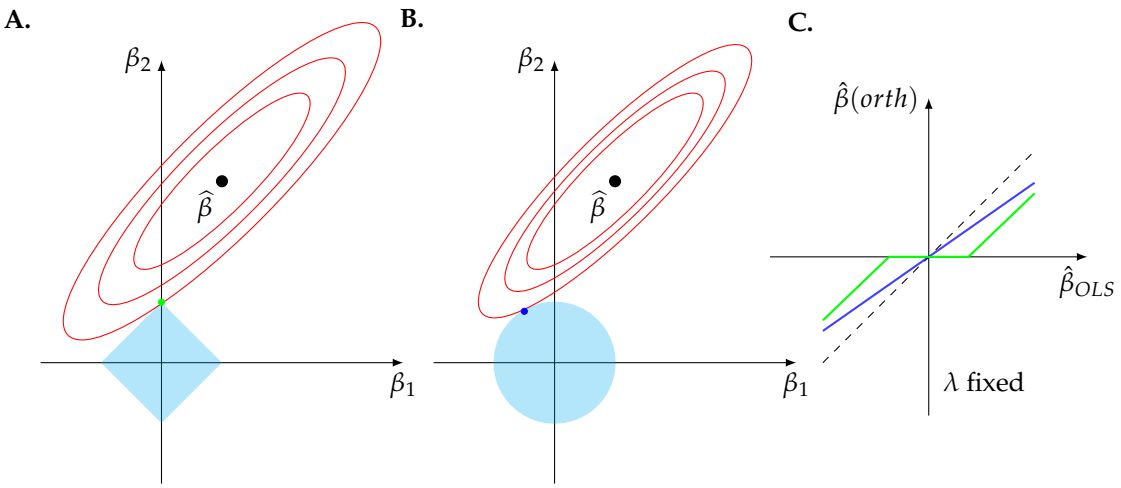

Figure 5. Visualization of the difference between the L1-norm (**A**) and L2-norm (**B**). (**C**) Solution for the orthonormal case.

In order to understand this, it is helpful to look at the solution for the coefficients for the orthonormal case, because for this situation, the solution for the LASSO can be found analytically. The analytical solution is given by

$$\hat{\beta}_i^{LASSO}(\lambda; orth) = \text{sign}(\beta_i^{\hat{OLS}}) S(\beta_i^{\hat{OLS}}, \lambda). \tag{26}$$

Here $S()$ is the soft-threshold operator defined as:

$$S(\beta_i^{\hat{OLS}}, \lambda) = \begin{cases} \beta_i^{\hat{OLS}} - \lambda & \text{if } \beta_i^{\hat{OLS}} > \lambda \\ 0 & \text{if } \left|\beta_i^{\hat{OLS}}\right| \leq \lambda \\ \beta_i^{\hat{OLS}} + \lambda & \text{if } \beta_i^{\hat{OLS}} < -\lambda \end{cases} \tag{27}$$

For the ridge regression the orthonormal solution is

$$\hat{\beta}_i^{RR}(\lambda; orth) = \frac{\beta_i^{\hat{OLS}}}{1+\lambda}. \tag{28}$$

In Figure 5C, we show Equation (26) (green) and Equation (28) (blue). As a reference, we added the ordinary least square solution as a dashed diagonal line (black) because it is just the identity mapping,

$$\hat{\beta}_i^{OLS}(orth) = \beta_i^{\hat{OLS}}. \tag{29}$$

As one can see, ridge regression leads to a change in the slope of the line and, hence, leads to a shrinkage of the coefficient. However, it does not lead to a zero coefficient except for the point in the origin of the coordinate system. In contrast, LASSO shrinks the coefficient to zero for $\left|\beta_i^{\hat{OLS}}\right| \leq \lambda$.

6.3. Discussion

The key idea of the LASSO is to realize that the theoretically ideal penalty to achieve sparsity is the L0-norm (i.e., $\|\boldsymbol{\beta}\|_0 = \#\text{non-zero elements}$, see Equation (6)), which is computationally intractable, but can be mimicked with the L1-norm which makes the optimization problem convex [33].

There are three major differences between ridge regression and the LASSO:

1. The non-differentiable corners of the L1-ball produce sparse models for sufficiently large values of λ.
2. The lack of rotational invariance limits the use of the singular value theory.
3. The LASSO has no analytic solution, making both computational and theoretical results more difficult.

The first point implies that the LASSO is better than OLS for the purpose of interpretation. With a large number of independent variables, we often would like to identify a smaller subset of these variables that exhibit the strongest effects. The sparsity of the LASSO is mainly counted as an advantage because it leads to a simpler interpretation.

6.4. Limitations

There is a number of limitations of the LASSO estimator, which causes problems for variable selection in certain situations.

1. In the $p > n$ case, the LASSO selects at most n variables. This could be a limiting factor if the true model consists of more than n variables.
2. The LASSO has no grouping property, that means it tends to select only one variable from a group of highly correlated variables.
3. In the $n > p$ case and high correlations between predictors, it has been observed that the prediction performance of the LASSO is inferior to the ridge regression.

6.5. Applications in the Literature

The LASSO has found ample applications to many different problems. For instance, in computational biology the LASSO has been used for analyzing gene expression data from mRNA and microRNA data [34,35] to address basic molecular biological questions. For studying diseases it has been used for investigating infection diseases [36], various cancer types [37,38], diabetes [39], and cardiovascular diseases [40]. In the computational social sciences the LASSO has been used to study data from social media [41], the stock market [42], economy [43], and political science [44]. Further application areas include robotics [45], climatology [46], and pharmacology [47].

In general, the widespread applications of the LASSO are due to the omnipresence of regression problems in essentially all areas of science. Also, the increasing availability of data in recent years outside the natural sciences, e.g., the social sciences or management, enabled this propagation.

6.6. R Package

An efficient implementation of the LASSO is available via the cyclical coordinate descent method by [48]. This method is accessible via the *glmnet* R package [28]. In [48] it was shown that regression models with thousands of repressors and samples can be estimated within seconds.

7. Bridge Regression

Bridge regression was suggested by Frank and Friedman [15]. It minimizes the RSS subject to a constraint depending on parameter q:

$$\hat{\boldsymbol{\beta}} = arg\min \left\{ \sum_{i=1}^{n} \left(y_i - \sum_j \beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\} \tag{30}$$

$$= arg\min \left\{ \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|^q \right\} \tag{31}$$

The regularization term has the form of Lq-norm, although q can assume all positive values, i.e., $q > 0$. For the special case $q = 2$, one obtains Ridge regression and for $q = 1$ the LASSO. Although, Bridge regression was introduced in 1993 before the LASSO, the model has not been studied at that time. This justifies the LASSO as a new method because [12] presented a full analysis.

8. Dantzig Selector

A regression model that was particularly introduced for the large p case ($p \gg n$) having many more parameters than observations is the Dantzig selector [14].

The regression model solves the following problem,

$$\hat{\boldsymbol{\beta}} = arg\min \left\{ \|\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\|_\infty + \lambda \|\boldsymbol{\beta}\|_1 \right\}. \tag{32}$$

Here the L∞ norm is the maximum absolute value of the components of the argument. It is worth remarking that in contrast to the LASSO, here a X^T is added to the Loss (residual sum) in Equation (32) to make the solution rotation-invariant.

8.1. Discussion

One advantage of the Dantzig selector is that it is computationally simple because, technically, it can be reduced to linear programming. This inspired the name of the method because George Dantzig did seminal work of the simplex method for linear programming [6]. As a consequence, this regression model can be used for even higher-dimensional data than the LASSO.

The disadvantages are similar to the LASSO except that it can result in more than n non-zero coefficients in the case $p > n$ [49]. Additionally, the Dantzig selector is sensitive to outliers because the L∞ norm is very sensitive to outliers. For practical application, the latter is of crucial importance.

8.2. Applications in the Literature

The Dantzig selector has been applied to a much lesser extend than the LASSO. However, some applications can be found for gene expression data [37,50].

8.3. R Package

A practical analysis for the Dantzig selector can be performed using the *flare* R package [51].

9. Adaptive LASSO

The adaptive LASSO has been introduced by [16] in order to have a LASSO model with oracle properties. An oracle procedure is one that has the following oracle properties:

- Consistency in variable selection
- Asymptotic normality

In simple terms the oracle property means that a model performs as well as if the true underlying model would be known [52]. Specifically, property one means that a model selects all non-zero coefficients with probability one, i.e., an oracle identifies the correct subset of true variables. Property two means that non-zero coefficients are estimated as if the true model would be known. It has been shown that the adaptive LASSO is an oracle procedure but the LASSO is not [16].

The basic idea of the adaptive LASSO is to introduce weights for the penalty for each regression coefficient. Specifically, the adaptive LASSO is a two-step procedure. In the first step a weight vector $\hat{\boldsymbol{w}}$ is estimated from OLS estimates of $\hat{\boldsymbol{\beta}_{init}}$ and a connection between both is given by

$$\hat{\boldsymbol{w}} = \frac{1}{|\hat{\boldsymbol{\beta}_{init}}|^{\gamma}}. \tag{33}$$

Here γ is again a tuning parameter that has to be positive, i.e., $\gamma > 0$.

Second, for this weight vector $\boldsymbol{w} = (w_1, \ldots, w_p)^T$ the following weighted LASSO is formulated by:

$$\hat{\boldsymbol{\beta}} = arg\min \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \sum_j \beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j| \right\} \tag{34}$$

$$= arg\min \left\{ \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j| \right\} \tag{35}$$

It can be shown that for certain data-dependent weight vectors, the adaptive LASSO has oracle properties. Frequent choices for $\hat{\boldsymbol{\beta}_{init}}$ are $\hat{\boldsymbol{\beta}_{init}} = \hat{\boldsymbol{\beta}_{OLS}}$ for the small p case ($p \ll n$) and $\hat{\boldsymbol{\beta}_{init}} = \hat{\boldsymbol{\beta}_{RR}}$ for the large p case ($p \gg n$).

The adaptive LASSO penalty can be seen as an approximation to the Lq penalties with $q = 1 - \gamma$. One advantage of the adaptive LASSO is that given appropriate initial estimates, the criterion Equation (34) is convex in β. Furthermore, if the initial estimates are N consistent, Zou (2006) showed that the method recovers the true model under more general conditions than the LASSO.

9.1. Example

In Figure 6 we show results for the economy data for $\gamma = 1$. In Figure 6A we show the coefficient paths in dependence on $\log(\lambda)$ and in Figure 6B the results for the mean-squared error. One can see the shrinking and selecting property of the adaptive LASSO because the regression coefficients become smaller for decreasing values of λ and some even vanish.

For the above results we used $\gamma = 1$, however, γ is a tuning parameter that can be estimated from the data. Specifically, in Figure 6C,D we repeated our analysis for different values of γ. From Figure 6C one can see that the minimal mean-squared error is obtained for $\gamma = 0.25$ but $\gamma = 1.0$ also gives good results. In Figure 6D we show the same results as in Figure 6C but for the mean-squared error in dependence on λ_{min}. There, one sees that the λ_{min} for large values of γ are quite large.

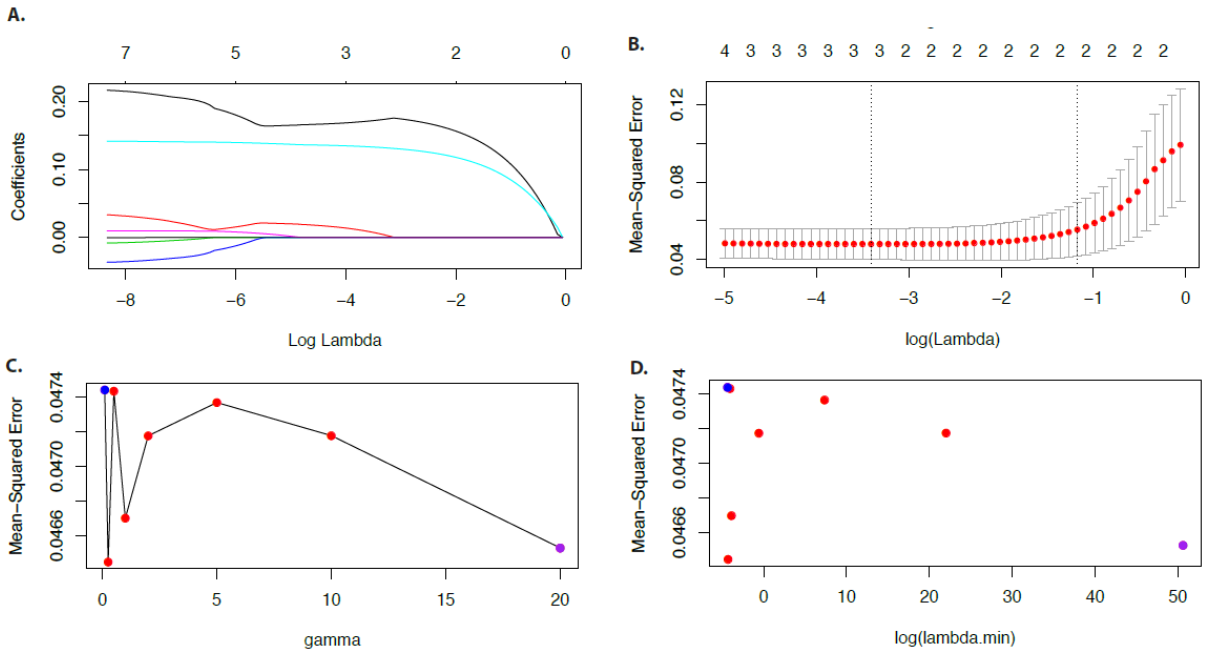

Figure 6. Adaptive LASSO model. (**A**) Coefficient paths in dependence on $\log(\lambda)$ for $\gamma = 1$. (**B**) Mean-squared error in dependence on $\log(\lambda)$. (**C**) Mean-squared error in dependence on γ. (**D**) Mean-squared error in dependence on $\log(\lambda_{min})$.

9.2. Applications in the Literature

The adaptive LASSO has also been applied to many different problems. For instance, in genomics the adaptive LASSO has been above all used to analyze quantitative trait loci (QTL) based on SNP (Single Nucleotide Polymorphism) measurements [7,53–55]. It has also been applied to analyze clinical data [10,56, 57] of various diseases, including cardiovascular and liver disease, and to assess organ transplantation [58].

9.3. R Package

A practical analysis for the adaptive LASSO can be performed using the *glmnet* R package [28].

10. Elastic Net

The elastic net regression model has been introduced by [17] to extend the LASSO by improving some of its limitations, especially with respect to the variable selection. Importantly, the elastic net encourages a grouping effect, keeping strongly correlated predictors together in the model. In contrast, the LASSO tends to split such groups keeping only the strongest variable. Furthermore, the elastic net is particularly useful in cases when the number of predictors (p) in a data set is much larger than the number of observations (n). In such a case, the LASSO is not capable of selecting more than n predictors but the elastic net has this capability.

Assuming standardized regressors and response, the elastic net solves the following problem:

$$\hat{\boldsymbol{\beta}} = arg\min\left\{\frac{1}{2N}\sum_{i=1}^{N}\left(y_i-\sum_j \beta_j x_{ij}\right)^2+\lambda P_\alpha(\beta)\right\} \quad (36)$$

$$= arg\min\left\{\frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\lambda P_\alpha(\beta)\right\} \quad (37)$$

$$P_\alpha(\beta) = \alpha\|\beta\|_2^2+(1-\alpha)\|\beta\|_1 \quad (38)$$

$$= \sum_{j=1}^{p}\alpha\beta_j^2+(1-\alpha)\left|\beta_j\right| \quad (39)$$

Here $P_\alpha(\beta)$ is the elastic net penalty (Zou and Hastie 2005). $P_\alpha(\beta)$ is a combination between the ridge regression penalty, for $\alpha = 1$, and the LASSO penalty, for $\alpha = 0$. This form of penalty turned out to be particularly useful in the case $p > n$, or in situations where we have many (highly) correlated predictor variables.

In the correlated case, it is known that ridge regression shrinks the regression coefficients of the correlated predictors towards each other. In the extreme case of k identical predictors, each of them obtains the same estimates of the coefficients [48]. From theoretical considerations it is further known that the ridge regression is optimal if there are many predictors, and all have non-zero coefficients. LASSO, on the other hand, is somewhat indifferent to very correlated predictors, and will tend to pick one and ignore the rest.

Interestingly, it is known that the elastic net with $\alpha = \epsilon$, for some very small $\epsilon > 0$, performs very similarly to the LASSO, but removes any degeneracies caused by the presence of correlations among the predictors [48]. More generally, the penalty family given by $P_\alpha(\beta)$ creates a non-trivial mixture between ridge regression and the LASSO. When for a given λ, one decreases α from 1 to 0, the number of regression coefficients equal to zero increases monotonically from 0 (full (ridge regression) model) to the sparsity of the LASSO solution. Here sparsity refers to the fraction of regression coefficients equal to zero. For more detail, see Friedman et al. [48] providing also an efficient implementation of the elastic net penalty for a variety of loss functions.

10.1. Example

In Figure 7 we show results for the economy data for $\alpha = 0.7$. Figure 7A shows the coefficient paths in dependence on $\log(\lambda)$ and Figure 7B the mean-squared error in dependence on $\log(\lambda)$.

Due to the fact that α is a parameter one needs to choose an optimal value. For this reason, we repeat the analysis for different values of α. Figure 7C,D show the results for the mean-squared error in dependence on α and the mean-squared error in dependence on $\log(\lambda_{\min})$. In these figures, $\alpha = 1.0$ corresponds to a ridge regression (blue point) and $\alpha = 0$ corresponds to the LASSO (purple point). As one can see, an α value of 0.7 leads to the minimal value of the mean-squared error and, hence, the optimal value of α.

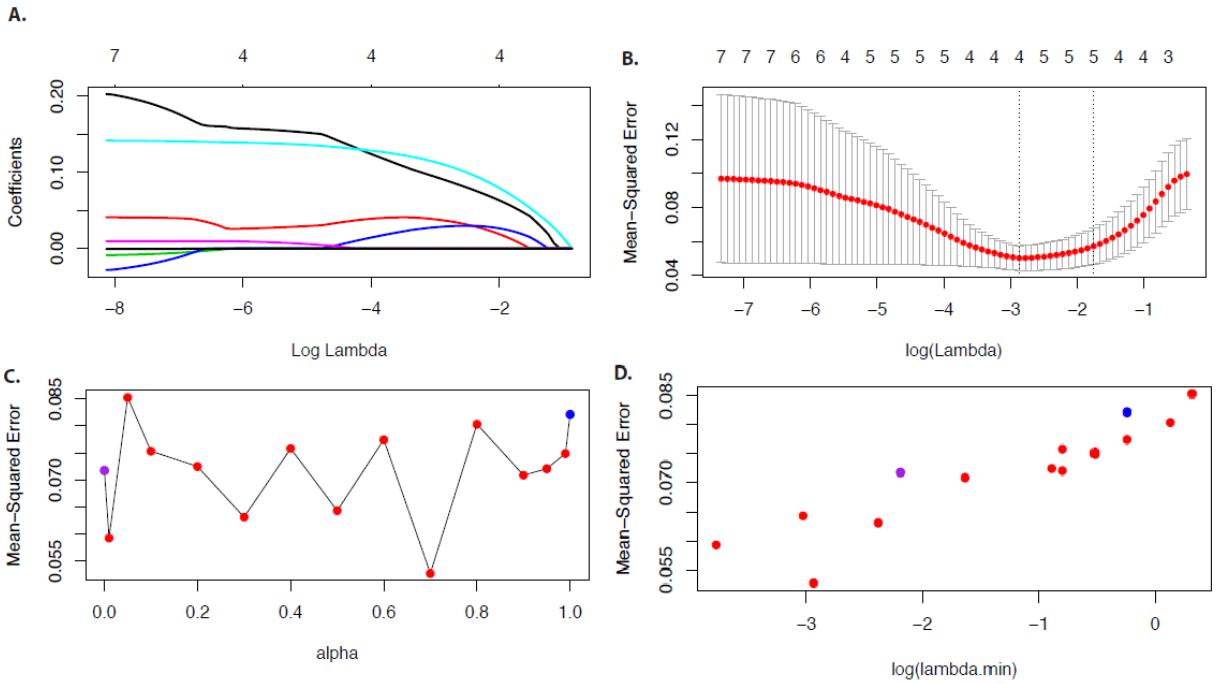

Figure 7. Elastic net. (**A**) Coefficient paths in dependence on $\log(\lambda)$ for $\alpha = 0.7$. (**B**) Mean-squared error in dependence on $\log(\lambda)$. (**C**) Mean-squared error in dependence on α. (**D**) Mean-squared error in dependence on $\log(\lambda_{min})$.

10.2. Discussion

The elastic net has been introduced to counteract the drawbacks of the LASSO and ridge regression. The ideas was to use a penalty for the elastic net which is based on a combined penalty of the LASSO and ridge regression. The penalty parameter α determines how much weight should be given to either the LASSO or ridge regression. An elastic net with $\alpha = 1.0$ performs a ridge regression and an elastic net with $\alpha = 0$ performs the LASSO. Specifically, several studies [59,60] showed that:

1. In the case of correlated predictors, the elastic net can result in lower mean squared errors compared to ridge regression and the LASSO.
2. In the case of correlated predictors, the elastic net selects all predictors whereas the LASSO selects one variable from a correlated group of variables but tends to ignore the remaining correlated variables.
3. In the case of uncorrelated predictors, the additional ridge penalty brings little improvement.
4. The elastic net identifies correctly a larger number of variables compared to the LASSO (model selection).
5. The elastic net has often a lower false positive rate compared to ridge regression.
6. In the case $p > n$, the elastic net can select more than n predictor variables whereas the LASSO selects at most n.

The last point means that the elastic net is capable of performing group selection of variables, at least to a certain degree. For further improving this property the group LASSO has been introduced (see below).

It can be shown that the elastic net penalty is a convex combination of the LASSO penalty and the ridge penalty. Specifically, for all $\alpha \in (0,1)$ the penalty function is strictly convex. In Figure 8, we visualize the effect of the tuning parameter α on the regularization. As one can see, the elastic net penalty (in red) is located between the LASSO penalty (in blue) and the ridge penalty (in green).

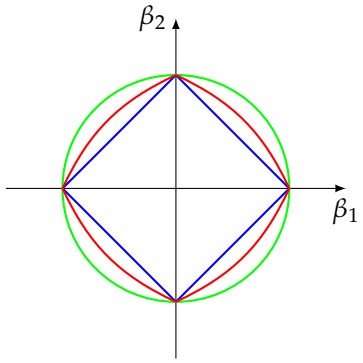

Figure 8. Visualization of the elastic net regularization (red) combining the L2-norm (green) of ridge regression and the L1-norm (blue) of LASSO.

The orthonormal solutions of the elastic net is similar to the LASSO in Equation (26). It is given by [17]

$$\hat{\beta}_i^{EN}(\lambda; orth) = \text{sign}(\beta_i^{\hat{OLS}}) \frac{S(\beta_i^{\hat{OLS}}, \lambda_1)}{1 + \lambda_2} \tag{40}$$

with $S(\beta_i^{\hat{OLS}}, \lambda_1)$ defined as:

$$S(\beta_i^{\hat{OLS}}, \lambda_1) = \begin{cases} \beta_i^{\hat{OLS}} - \lambda_1 & \text{if } \beta_i^{\hat{OLS}} > \lambda_1 \\ 0 & \text{if } \left|\beta_i^{\hat{OLS}}\right| \leq \lambda_1 \\ \beta_i^{\hat{OLS}} + \lambda_1 & \text{if } \beta_i^{\hat{OLS}} < -\lambda_1 \end{cases} \tag{41}$$

Here the parameters λ_1 and λ_2 are connected to λ and α in Equation (36) by

$$\alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2} \tag{42}$$

$$\lambda = \lambda_1 + \lambda_2 \tag{43}$$

resulting in the alternative form of the elastic net

$$\hat{\boldsymbol{\beta}} = arg\min \left\{ \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1 \right\}. \tag{44}$$

In contract to the LASSO in Equation (26), only the slope of the line for $\left|\beta_i^{\hat{OLS}}\right| > \lambda_1$ is different due to the denominator $1 + \lambda_2$. That means the ridge penalty, controlled by λ_2, performs a second shrinkage effect on the coefficients. Hence, an elastic net performs a *double shrinkage* on the coefficients, one from the LASSO penalty and one from the ridge penalty. Hence, from Equation (40) one can also see the variable selection property of the elastic net, similar to the LASSO.

10.3. Applications in the Literature

Due to the improved characteristics of the elastic net over the LASSO, this method is frequently preferred. For instance, in genomics it has been used for genome-wide association studies (GWAS) studying SNPs [7,61–63]. Gene expression data have also been studied, e.g., to identify prognostic biomarkers for

breast cancer [64] or for drug repurposing [65]. Furthermore, electronic patient health records have been analyzed for predicting patient mortality [66]. In imaging, resting state functional magnetic resonance imaging (RSfMRI) was studied to identify patients with Alzheimer's disease [67]. In finance, elastic nets have been used to define portfolios of stocks [68] or to predict the credit ratings of corporations [69].

10.4. R Package

A practical analysis of the elastic net can be performed using the *glmnet* R package [28].

11. Group LASSO

The last modern regression model we are discussing is the group LASSO, introduced by [18]. The group LASSO is different to the other regression models because it focuses on groups of variables instead of individual variables. The reason for this is that there are many real-world application problems related to, e.g., pathways of genes, portfolios of stocks, or substage disorders of patients, which have substructures, whereas a set of predictors forms a group that either should have nonzero or zero coefficients simultaneously.

The various forms of group lasso penalty are designed for such situations.

Let us suppose that the p predictors are divided into G groups, and p_g is the number of predictors in group $g \in \{1, \ldots, G\}$. The matrix $\boldsymbol{X}_g \in \mathbb{R}^{n \times p_g}$ represents the predictors corresponding to group g and the corresponding regression coefficient vector is given by $\boldsymbol{\beta}_g \in \mathbb{R}^{p_g}$.

The group LASSO solves the following convex optimization problem:

$$\hat{\boldsymbol{\beta}} \quad = \quad arg\min \left\{ \frac{1}{2n} \|\boldsymbol{y} - \sum_{g=1}^{G} \boldsymbol{X}_g \boldsymbol{\beta}_g\|_2^2 + \lambda \sum_{g=1}^{G} \sqrt{p_g} \|\boldsymbol{\beta}_g\|_2 \right\} \tag{45}$$

Here the term p_g accounts for the varying group sizes. If $p_g = 1$ for all groups g, then the group LASSO becomes the ordinary LASSO. If $p_g > 1$, the group LASSO works like the LASSO but on the group level, instead of the individual predictors.

11.1. Example

In Figure 9 we show results for the economy data for the group labels $\{1, 1, 1, 2, 2, 3, 3\}$ for the seven predictors. Figure 9A shows the coefficient paths in dependence on $\log(\lambda)$ and Figure 9B the mean-squared error in dependence on $\log(\lambda)$. From the number of variables above each figure, one can see that 1 and 3 never appear. The former is for principle reasons not possible because the smallest group (labeled by '2' or '3') consists of two predictors and the latter does just not occur in this example because the group labeled by '1' (consisting of three predictors) is added *later* to the model for larger λ values. Hence, there are jumps in the number of predictors.

Due to the fact that for this data set no obvious grouping of the predictors is available, we repeat the above analysis for 640 different group definitions for two to three different groups. In Figure 9C we show the results for this analysis. The y-axis shows the minimum mean-squared errors of the corresponding models corresponding to λ_{min}. The x-axis enumerates these models from 1 to 640 and the legend gives the color code for the optimal number of variables that minimize the MSE.

The last analysis demonstrates also a problem of the group LASSO because if the group definitions of the predictors are not known or cannot be obtained in a natural way, e.g., by interpretation of the problem under investigation, searching for the optimal grouping of the predictors becomes, even for a relatively small number of variables, computationally demanding due to the large number of possible combinations.

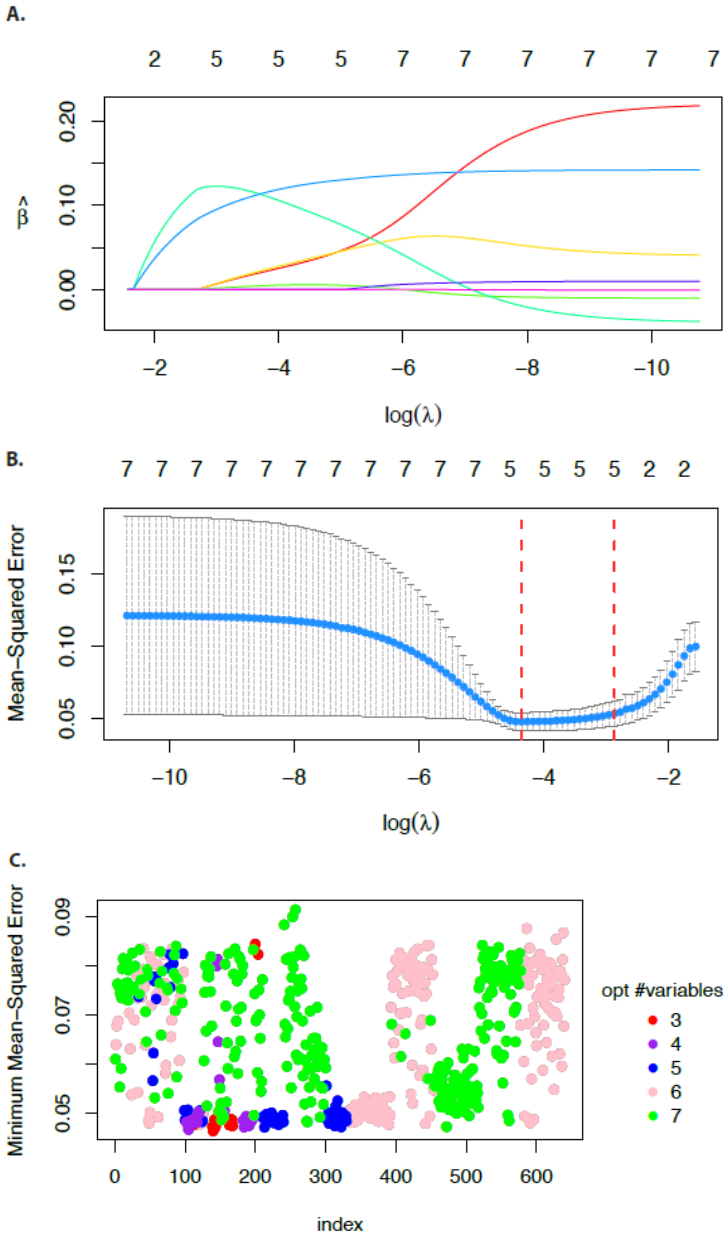

Figure 9. Group LASSO. (**A**) Coefficient paths in dependence on $\log(\lambda)$. (**B**) Mean-squared error in dependence on $\log(\lambda)$. (**C**) Minimum Mean-squared error for 640 different group definitions. The legend gives the color code for the optimal number of variables that minimize the MSE.

11.2. Discussion

1. The group LASSO has either zero coefficients of all members of a group or non-zero coefficients.
2. The group LASSO cannot achieve sparsity within a group.
3. The groups need to be predefined, i.e., the regression model does not provide a direct mechanism to obtain the grouping.
4. The groups are mutually exclusive (non-overlapping).

Finally, we just want to briefly mention that to overcome the limitation of the group LASSO to obtain sparsity within a group (point (2)), the sparse group LASSO has been introduced by [70]. The model is defined by:

$$\hat{\boldsymbol{\beta}} \quad = \quad arg\min \left\{ \frac{1}{2n} \|\boldsymbol{y} - \sum_{g=1}^{G} \boldsymbol{X}_g \boldsymbol{\beta}_g\|_2^2 + (1-\alpha)\lambda \sum_{g=1}^{G} \sqrt{p_g} \|\boldsymbol{\beta}_g\|_2 + \alpha\lambda \|\boldsymbol{\beta}\|_1 \right\} \tag{46}$$

For $\alpha \in [0,1]$ this is a convex optimization problem combining the group LASSO penalty (for $\alpha = 0$) with the LASSO penalty (for $\alpha = 1$). Here $\boldsymbol{\beta} \in \mathbb{R}^p$ is the complete coefficient vector.

11.3. Applications in the Literature

Examples for applications of the group LASSO can be found above all in genomics where groups can be naturally defined via biological processes to which genes are belonging. For instance, the group LASSO has been applied to GWAS data [71,72] and gene expression data [70,73].

11.4. R Package

A practical analysis of the group LASSO can be performed using the *oem* R package [74,75]. Also, the *gglasso* R package performs a group LASSO [76].

12. Summary

In this paper, we surveyed modern regression models that extend OLS regression. In contrast to the OLS regression and ridge regression, all of these models are computational in nature because the solution to the various regularizations can only be found by means of numerical approaches.

In Table 1, we summarize some key features of these regression models. A common feature of all extensions of OLS regression and ridge regression is that these models perform variable selection (coefficient shrinkage to zero). This allows to obtain interpretable models because the smaller the number of variables in a model, the easier it is to find plausible explanations. Considering this, the adaptive LASSO has the most satisfying properties because it possesses the oracle property, making it capable to identify only the coefficients that are non-zero in the true model.

Table 1. Summary of key features of the regression models.

Method	Analytical Solution	Variable Selection	Can Select $> n$	Grouping	Oracle
ridge regression	yes	no	yes	yes	no
non-negative garotte	no	yes	no	no	no
LASSO	no	yes	no	no	no
Dantzig selector	no	yes	yes	no	no
adaptive LASSO	no	yes	no	no	yes
elestic net	no	yes	yes	yes	no
group LASSO	no	yes	yes	yes	no

In general, one considers data as high-dimensional if either (I) p is large or (II) $p > n$ [59,77,78]. The case (I) can be handled by all regression models, including the OLS regression. However, case (II) is more difficult because it may require to select more variables than samples are available. Only ridge regression, Dantzig selector, elastic net, and the group LASSO are capable of this, and the elastic net is particularly suited for this situation.

Finally, the grouping of variables is useful, e.g., in cases when variables are highly correlated with each other. Again ridge regression, the elastic net, and the group LASSO have this property, and the group LASSO has been specifically introduced to deal with this problem.

In Table 2, we summarize the regularization terms of the different models. From this one can understand why the number of new regularized regression models exploded in recent years because these models explore different forms of the Lq-norm or combine different norms with each other. This proved a very rich source of inspiration and new models are still under development.

We did not include the non-negative garotte and the Dantzig selector in this table because these models have a different form of the RSS term.

Table 2. Overview of regularization or penalty terms and methods utilizing them.

Regularization Term	Method
L2 norm: $\|\boldsymbol{\beta}\|_2$	ridge regression
L1 norm: $\|\boldsymbol{\beta}\|_1$	LASSO
Lq norm: $\|\boldsymbol{\beta}\|_q$	bridge regression
weighted L1 norm: $\sum_{j=1}^{p} w_j \left\|\beta_j\right\|$	adaptive LASSO
$c_1\|\boldsymbol{\beta}\|_1 + c_2\|\boldsymbol{\beta}\|_2^2$	elastic net
$\sum_{g=1}^{G} \sqrt{p_g}\|\beta_g\|_2$	group LASSO

Regarding practical applications of these regularization regression models, the following is important to note:

> There is not one model that is always and under all conditions better than all other models. Instead, the performance of all of these models are highly *data-dependent*.

Specifically, the characteristics of a data set have a strong influence on the performance of a regression model. For this reason, for practical applications it is strongly advisable to perform a comparative analysis of different models for a particular data set under investigation by means of cross validation (CV), potentially in combination with simulation studies that mimic the characteristics within this data set. Furthermore, it would be beneficial to analyze more than one data set (validation data) of the same data type to obtain reliable estimates of the variabilities among the covariates. Only in this way is it possible to guard against false assumptions leading to the selection of a suboptimal model for the data set under investigation.

13. Conclusions

Regression models find widespread applications in science and our digital society. Over the years, many different regularization models have been introduced, where each addresses a particular problem, making no one method dominant over the others, since they all have specific strengths and weaknesses. The LASSO and related models are very popular tools in this context that form core methods of modern data science [79]. Despite this, there is a remarkable lack in the literature regarding accessible reviews on the intermediate level. Our review aims to fill this gap, with a particular focus on the regularization terms.

Author Contributions: F.E.-S. conceived the study. All authors contributed to the writing of the manuscript and approved the final version.

Funding: M.D. thanks the Austrian Science Funds for supporting this work (project P30031).

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Chang, R.M.; Kauffman, R.; Kwon, Y. Understanding the paradigm shift to computational social science in the presence of big data. *Decis. Support Syst.* **2014**, *63*, 67–80. [CrossRef]
2. Emmert-Streib, F.; Yli-Harja, O.; Dehmer, M. Data analytics applications for streaming data from social media: What to predict? *Front. Big Data* **2018**, *1*, 1. [CrossRef]
3. Dehmer, M.; Emmert-Streib, F.; Graber, A.; Salvador, A. (Eds.) *Applied Statistics for Network Biology: Methods for Systems Biology*; Wiley-Blackwell: Weinheim, Germany, 2011.
4. Emmert-Streib, F.; Altay, G. Local network-based measures to assess the inferability of different regulatory networks. *IET Syst. Biol.* **2010**, *4*, 277–288. [CrossRef] [PubMed]
5. Harrell, F.E. *Regression Modeling Strategies*; Springer: New York, NY, USA, 2001.
6. Haste, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*; Springer: New York, NY, USA, 2009.
7. Ogutu, J.O.; Schulz-Streeck, T.; Piepho, H.P. Genomic selection using regularized linear regression models: Ridge regression, lasso, elastic net and their extensions. *BMC Proc.* **2012**, *6*, S10. [CrossRef] [PubMed]
8. Emmert-Streib, F.; Dehmer, M. Defining Data Science by a Data-Driven Quantification of the Community. *Mach. Learn. Knowl. Extr.* **2019**, *1*, 235–251. [CrossRef]
9. Li, Z.; Sillanpää, M.J. Overview of LASSO-related penalized regression methods for quantitative trait mapping and genomic selection. *Theor. Appl. Genet.* **2012**, *125*, 419–435. [CrossRef] [PubMed]
10. Lu, M.; Zhou, J.; Naylor, C.; Kirkpatrick, B.D.; Haque, R.; Petri, W.A.; Ma, J.Z. Application of penalized linear regression methods to the selection of environmental enteropathy biomarkers. *Biomark. Res.* **2017**, *5*, 9. [CrossRef]
11. Yeung, R. *A First Course in Information Theory*; Springer: Berlin/Heidelberg, Germany, 2002.
12. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1996**, *58*, 267–288. [CrossRef]
13. Breiman, L. Better subset regression using the nonnegative garrote. *Technometrics* **1995**, *37*, 373–384. [CrossRef]
14. Candes, E.; Tao, T. The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Stat.* **2007**, *35*, 2313–2351. [CrossRef]
15. Frank, L.E.; Friedman, J.H. A statistical view of some chemometrics regression tools. *Technometrics* **1993**, *35*, 109–135. [CrossRef]
16. Zou, H. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **2006**, *101*, 1418–1429. [CrossRef]
17. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2005**, *67*, 301–320. [CrossRef]
18. Yuan, M.; Lin, Y. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2006**, *68*, 49–67. [CrossRef]
19. Dasgupta, A.; Sun, Y.V.; König, I.R.; Bailey-Wilson, J.E.; Malley, J.D. Brief review of regression-based and machine learning methods in genetic epidemiology: The Genetic Analysis Workshop 17 experience. *Genet. Epidemiol.* **2011**, *35*, S5–S11. [CrossRef] [PubMed]
20. Huang, J.; Breheny, P.; Ma, S. A selective review of group selection in high-dimensional models. *Stat. Sci. Rev. J. Inst. Math. Stat.* **2012**, *27*. [CrossRef]
21. Song, Q. An overview of reciprocal L 1-regularization for high dimensional regression data. *Wiley Interdiscip. Rev. Comput. Stat.* **2018**, *10*, e1416. [CrossRef]
22. Tikhonov, A.N. On the stability of inverse problems. *Dokl. Akad. Nauk SSSR* **1943**, *39*, 195–198.
23. Bickel, P.J.; Li, B. Regularization in statistics. *Test* **2006**, *15*, 271–344. [CrossRef]

24. Garcia, M.G.; Medeiros, M.C.; Vasconcelos, G.F. Real-time inflation forecasting with high-dimensional models: The case of Brazil. *Int. J. Forecast.* **2017**, *33*, 679–693. [CrossRef]
25. Kaufman, R.L. *Heteroskedasticity in Regression: Detection and Correction*; Sage Publications: Thousand Oaks, CA, USA, 2013; Volume 172.
26. Hastie, T.; Tibshirani, R.; Wainwright, M. *Statistical Learning with Sparsity: The Lasso and Generalizations*; CRC Press: Boca Raton, FL, USA, 2015.
27. Hoerl, A.E.; Kennard, R.W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **1970**, *12*, 55–67. [CrossRef]
28. Friedman, J.; Hastie, T.; Tibshirani, R. *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models*, R Package Version 2009. Available online: https://cran.r-project.org/web/packages/glmnet/index.html (accessed on 9 December 2018).
29. Yuan, M.; Lin, Y. On the non-negative garrotte estimator. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2007**, *69*, 143–161. [CrossRef]
30. Fan, J.; Lv, J. A selective overview of variable selection in high dimensional feature space. *Stat. Sin.* **2010**, *20*, 101. [PubMed]
31. Santosa, F.; Symes, W.W. Linear inversion of band-limited reflection seismograms. *SIAM J. Sci. Stat. Comput.* **1986**, *7*, 1307–1330. [CrossRef]
32. Zou, H.; Hastie, T.; Tibshirani, R. On the "degrees of freedom" of the lasso. *Ann. Stat.* **2007**, *35*, 2173–2192. [CrossRef]
33. Van de Geer, S. L1-regularization in High-dimensional Statistical Models. In Proceedings of the International Congress of Mathematicians 2010 (ICM 2010), Hyderabad, India, 19–27 August 2010; pp. 2351–2369. [CrossRef]
34. Cosgrove, E.J.; Zhou, Y.; Gardner, T.S.; Kolaczyk, E.D. Predicting gene targets of perturbations via network-based filtering of mRNA expression compendia. *Bioinformatics* **2008**, *24*, 2482–2490. [CrossRef]
35. Lu, Y.; Zhou, Y.; Qu, W.; Deng, M.; Zhang, C. A Lasso regression model for the construction of microRNA- target regulatory networks. *Bioinformatics* **2011**, *27*, 2406–2413. [CrossRef]
36. Chen, Y.; Chu, C.W.; Chen, M.I.; Cook, A.R. The utility of LASSO-based models for real time forecasts of endemic infectious diseases: A cross country comparison. *J. Biomed. Inform.* **2018**, *81*, 16–30. [CrossRef]
37. Zheng, S.; Liu, W. An experimental comparison of gene selection by Lasso and Dantzig selector for cancer classification. *Comput. Biol. Med.* **2011**, *41*, 1033–1040. [CrossRef]
38. Daniels, M.W.; Brock, G.N.; Wittliff, J.L. Clinical outcomes linked to expression of gene subsets for protein hormones and their cognate receptors from LCM-procured breast carcinoma cells. *Breast Cancer Res. Treat.* **2017**, *161*, 245–258. [CrossRef]
39. Nowak, C.; Sundström, J.; Gustafsson, S.; Giedraitis, V.; Lind, L.; Ingelsson, E.; Fall, T. Protein biomarkers for insulin resistance and type 2 diabetes risk in two large community cohorts. *Diabetes* **2015**, *65*, 276–284. [CrossRef] [PubMed]
40. You, M.; Fang, W.; Wang, X.; Yang, T. Modelling of the ICF core sets for chronic ischemic heart disease using the LASSO model in Chinese patients. *Health Qual. Life Outcomes* **2018**, *16*, 139. [CrossRef] [PubMed]
41. Bovet, A.; Morone, F.; Makse, H.A. Validation of Twitter opinion trends with national polling aggregates: Hillary Clinton vs Donald Trump. *Sci. Rep.* **2018**, *8*, 8673. [CrossRef] [PubMed]
42. Roy, S.S.; Mittal, D.; Basu, A.; Abraham, A. Stock market forecasting using LASSO linear regression model. In *Afro-European Conference for Industrial Advancement*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 371–381.
43. Tian, S.; Yu, Y.; Guo, H. Variable selection and corporate bankruptcy forecasts. *J. Bank. Finance* **2015**, *52*, 89–100. [CrossRef]
44. Mauerer, I.; Pößnecker, W.; Thurner, P.W.; Tutz, G. Modeling electoral choices in multiparty systems with high-dimensional data: A regularized selection of parameters using the lasso approach. *J. Choice Model.* **2015**, *16*, 23–42. [CrossRef]
45. Do, H.N.; Choi, J.; Lim, C.Y.; Maiti, T. Appearance-Based Localization of Mobile Robots Using Group LASSO Regression. *J. Dyn. Syst. Meas. Control* **2018**, *140*, 091016. [CrossRef]

46. Tan, J.; Liu, H.; Li, M.; Wang, J. A prediction scheme of tropical cyclone frequency based on lasso and random forest. *Theor. Appl. Climatol.* **2018**, *133*, 973–983. [CrossRef]
47. Ahmed, I.; Pariente, A.; Tubert-Bitter, P. Class-imbalanced subsampling lasso algorithm for discovering adverse drug reactions. *Stat. Methods Med. Res.* **2018**, *27*, 785–797. [CrossRef]
48. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **2010**, *33*, 1. [CrossRef]
49. Efron, B.; Hastie, T.; Tibshirani, R. Discussion: The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Stat.* **2007**, *35*, 2358–2364. [CrossRef]
50. Vignes, M.; Vandel, J.; Allouche, D.; Ramadan-Alban, N.; Cierco-Ayrolles, C.; Schiex, T.; Mangin, B.; De Givry, S. Gene regulatory network reconstruction using Bayesian networks, the Dantzig Selector, the Lasso and their meta-analysis. *PLoS ONE* **2011**, *6*, e29165. [CrossRef] [PubMed]
51. Li, X.; Zhao, T.; Yuan, X.; Liu, H. The flare package for high dimensional linear regression and precision matrix estimation in R. *J. Mach. Learn. Res.* **2015**, *16*, 553–557. [PubMed]
52. Zhou, N.; Zhu, J. Group variable selection via a hierarchical lasso and its oracle property. *arXiv* **2010**, arXiv:1006.2871.
53. Sun, W.; Ibriham, J.G.; Zou, F. Genome-wide multiple loci mapping in experimental crosses by the iterative adaptive penalized regression. *Genetics* **2010**. [CrossRef] [PubMed]
54. Li, M.; Romero, R.; Fu, W.J.; Cui, Y. Mapping haplotype-haplotype interactions with adaptive LASSO. *BMC Genet.* **2010**, *11*, 79. [CrossRef] [PubMed]
55. He, Q.; Lin, D.Y. A variable selection method for genome-wide association studies. *Bioinformatics* **2010**, *27*, 1–8. [CrossRef] [PubMed]
56. Dai, L.; Koutrakis, P.; Coull, B.A.; Sparrow, D.; Vokonas, P.S.; Schwartz, J.D. Use of the adaptive LASSO method to identify PM2.5 components associated with blood pressure in elderly men: The Veterans Affairs Normative Aging Study. *Environ. Health Perspect.* **2015**, *124*, 120–125. [CrossRef] [PubMed]
57. Corey, K.E.; Kartoun, U.; Zheng, H.; Shaw, S.Y. Development and validation of an algorithm to identify nonalcoholic fatty liver disease in the electronic medical record. *Dig. Dis. Sci.* **2016**, *61*, 913–919. [CrossRef]
58. Raeisi Shahraki, H.; Pourahmad, S.; Ayatollahi, S.M.T. Identifying the prognosis factors in death after liver transplantation via adaptive LASSO in Iran. *J. Environ. Public Health* **2016**, *2016*, 1–6. [CrossRef]
59. Bühlmann, P.; Van De Geer, S. *Statistics for High-Dimensional Data: Methods, Theory and Applications*; Springer: Berlin/Heidelberg, Germany, 2011.
60. Tutz, G.; Ulbricht, J. Penalized regression with correlation-based penalty. *Stat. Comput.* **2009**, *19*, 239–253. [CrossRef]
61. Waldmann, P.; Mészáros, G.; Gredler, B.; Fuerst, C.; Sölkner, J. Evaluation of the lasso and the elastic net in genome-wide association studies. *Front. Genet.* **2013**, *4*, 270. [CrossRef] [PubMed]
62. Momen, M.; Mehrgardi, A.A.; Sheikhi, A.; Kranis, A.; Tusell, L.; Morota, G.; Rosa, G.J.; Gianola, D. Predictive ability of genome-assisted statistical models under various forms of gene action. *Sci. Rep.* **2018**, *8*, 12309. [CrossRef] [PubMed]
63. Lu, Y.; Beeghly-Fadiel, A.; Wu, L.; Guo, X.; Li, B.; Schildkraut, J.M.; Im, H.K.; Chen, Y.A.; Permuth, J.B.; Reid, B.M.; et al. A transcriptome-wide association study among 97,898 women to identify candidate susceptibility genes for epithelial ovarian cancer risk. *Cancer Res.* **2018**, *78*, 5419–5430. [CrossRef] [PubMed]
64. Buus, R.; Yeo, B.; Brentnall, A.R.; Klintman, M.; Cheang, M.C.U.; Khabra, K.; Sestak, I.; Gao, Q.; Cuzick, J.; Dowsett, M. Novel 18-gene signature for predicting relapse in ER-positive, HER2-negative breast cancer. *Breast Cancer Res.* **2018**, *20*, 103. [CrossRef] [PubMed]
65. Wang, Y.; Wang, Z.; Xu, J.; Li, J.; Li, S.; Zhang, M.; Yang, D. Systematic identification of non-coding pharmacogenomic landscape in cancer. *Nat. Commun.* **2018**, *9*, 3192. [CrossRef] [PubMed]
66. Steele, A.J.; Cakiroglu, S.A.; Shah, A.D.; Denaxas, S.C.; Hemingway, H.; Luscombe, N.M. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *bioRxiv* **2018**, 256008. [CrossRef]

67. De Vos, F.; Koini, M.; Schouten, T.M.; Seiler, S.; van der Grond, J.; Lechner, A.; Schmidt, R.; de Rooij, M.; Rombouts, S.A. A comprehensive analysis of resting state fMRI measures to classify individual patients with Alzheimer's disease. *NeuroImage* **2018**, *167*, 62–72. [CrossRef]
68. Ho, M.; Sun, Z.; Xin, J. Weighted elastic net penalized mean-variance portfolio design and computation. *SIAM J. Financ. Math.* **2015**, *6*, 1220–1244. [CrossRef]
69. Sermpinis, G.; Tsoukas, S.; Zhang, P. Modelling market implied ratings using LASSO variable selection techniques. *J. Empir. Finance* **2018**, *48*, 19–35. [CrossRef]
70. Simon, N.; Friedman, J.; Hastie, T.; Tibshirani, R. A sparse-group lasso. *J. Comput. Graph. Stat.* **2013**, *22*, 231–245. [CrossRef]
71. Chen, L.S.; Hutter, C.M.; Potter, J.D.; Liu, Y.; Prentice, R.L.; Peters, U.; Hsu, L. Insights into colon cancer etiology via a regularized approach to gene set analysis of GWAS data. *Am. J. Hum. Genet.* **2010**, *86*, 860–871. [CrossRef] [PubMed]
72. Ogutu, J.O.; Piepho, H.P. Regularized group regression methods for genomic prediction: Bridge, MCP, SCAD, group bridge, group lasso, sparse group lasso, group MCP and group SCAD. *BMC Proc.* **2014**, *8*, S7. [CrossRef] [PubMed]
73. Ma, S.; Song, X.; Huang, J. Supervised group Lasso with applications to microarray data analysis. *BMC Bioinform.* **2007**, *8*, 60. [CrossRef] [PubMed]
74. Xiong, S.; Dai, B.; Huling, J.; Qian, P.Z.G. Orthogonalizing EM: A design-based least squares algorithm. *Technometrics* **2016**, *58*, 285–293. [CrossRef] [PubMed]
75. Huling, J.D.; Chien, P. Fast Penalized Regression and Cross Validation for Tall Data with the oem Package. *J. Stat. Softw.* **2018**, arXiv:1801.09661.
76. Yang, Y.; Zou, H. *gglasso: Group Lasso Penalized Learning Using a Unified BMD Algorithm*, R Package Version 2014. Available online: https://cran.r-project.org/web/packages/gglasso/index.html (accessed on 9 December 2018).
77. Johnstone, I.M.; Titterington, D.M. Statistical challenges of high-dimensional data. *Philos. Trans. Ser. A Math. Phys. Eng. Sci.* **2009**, *367*, 4237. [CrossRef] [PubMed]
78. Meinshausen, N.; Yu, B. Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Stat.* **2009**, *37*, 246–270. [CrossRef]
79. Emmert-Streib, F.; Moutari, S.; Dehmer, M. The process of analyzing data is the emergent feature of data science. *Front. Genet.* **2016**, *7*, 12. [CrossRef]