ALBERT-LUDWIGS-
UNIVERSITÄT FREIBURG

information
systems

Business Analytics

– SEMINAR SUMMER SEMESTER 2014 –

Ridge Regression and Lasso

– SEMINAR PAPER –

Submitted by:

Student-ID:

Advisor:
Prof. Dr. Dirk Neumann

Contents

1 Introduction

Linear regression is a supervised learning model used for prediction and for explaining the relationship between a dependent and an independent variable [5]. Ordinary Least Squares (OLS) is the most commonly used method for fitting the linear regression model. Generally, the main purpose of a model is to use some data in order to predict future observations. Prediction accuracy and model complexity are two important concepts to be taken into account when performing prediction or when explaining relationships between variables [1]. The OLS method displays low performance with respect to these two criteria. OLS exhibit several statistical and numerical problems which result in variable coefficient estimates, low predictive power and not easily interpretable results. Some of the problems include multicollinearity, a number of predictors higher than the number of observations or high variance.

Regularization techniques such as lasso and ridge regression overcome some of the problems of OLS. Ridge regression shrinks the coefficients estimates towards zero, in this way improving the variability, reducing the variance while slightly increasing the bias and raising the overall accuracy of the model. However, when the number of predictors is high, ridge regression results in models that are difficult to interpret. Another regularization technique called *lasso* improves both accuracy and model interpretability by selecting which coefficients to shrink and by shrinking some of them exactly to zero [7].

This paper provides an introduction into these two techniques. It is structured as follows. As both of these technique extend the basic OLS model, Section 2 starts with a brief theoretical background of OLS and discusses some of its limitations. To overcome some of the limitations of OLS, we introduce ridge regression in Section 3. We start by motivating ridge regression with the bias-variance trade-off, then introduce its mathematical formulation and finally discuss how this method performs shrinkage. As an improvement of the ridge technique, we present the lasso in Section 4. Like in the case of ridge, we cover its mathematical formulation and the shrinkage method. After understanding the basic principles of both methods, we look at them comparatively through a geometrical and Bayesian lens in Section 5. Finally, we conclude in Section 6.

2 Ordinary Least Squares

Both ridge regression and lasso build upon OLS. We first look at this model and its limitations in this chapter.

2.1 Estimation of Ordinary Least Squares

Given the n tuples of training data $(y_i, x_{1\mathrm{i}}, \ldots, x_{\mathrm{ni}})$, a multiple regression model describing the relationship between the dependent variable y and the independent variables x_j is

$$y = \beta_0 + \sum_{j=1}^{N} \beta_j x_j + \varepsilon, \tag{1}$$

where ε is the error term and β_0 the intercept. We predict the value of y based on the value of x_j, by computing the following equation

$$\hat{y} = \hat{\beta}_0 + \sum_{j=1}^{N} \hat{\beta}_j x_j, \tag{2}$$

where the hat symbol indicates the estimated coefficients. In order to fit this model, we typically use a method called *least squares* or *Ordinary Least Squares (OLS)*. The idea behind OLS is to find the parameters $\hat{\beta}_j$ such that the regression line is the closest line to the data points (x_{ni}, y) [8]. The OLS method estimates the coefficients $\beta_1, \ldots, \beta_{\mathrm{n}}$ by minimizing a quantity known as the *Residual Sum of Squares (RSS)*. This is defined as

$$\mathrm{RSS} = \sum_{j=1}^{N} (y_i - \hat{y})^2 = \sum_{j=1}^{N} [y_i - (\hat{\beta}_0 + \sum_{j=1}^{N} \hat{\beta}_j x_j)]^2, \tag{3}$$

where y_i is the actual value and $\hat{y}$ is the predicted value. Differentiating the equation above with respect to $\hat{\beta}_j$ and solving for $\hat{\beta}_j$ yields the OLS coefficients, which we denote $\hat{\beta}_{\mathrm{OLS}}$. Hence, the OLS estimates are

$$\hat{\beta}_{\mathrm{OLS}} = \underset{\beta}{\operatorname{argmin}} \sum_{j=1}^{N} [y_i - (\beta_0 + \sum_{j=1}^{N} \beta_j x_j)]^2. \tag{4}$$

In order to gain more insight into how OLS work, we express the coefficients estimates in matrix algebra representation. In this case, the regression model is $y = X\beta + u$, where y is the output vector of n observations, β is a $(p+1) \times 1$ vector and X is a $n \times (p+1)$ matrix with 1's on the first column and each row an input vector. Therefore, the *RSS* expression from Equation (3) becomes

$$\mathrm{RSS} = (y - X\beta)^T (y - X\beta). \tag{5}$$

By differentiating this quantity and performing some calculations, we obtain the OLS coefficient estimates in matrix form

$$\hat{\beta}_{\mathrm{OLS}} = (X^T X)^{-1} X^T y, \tag{6}$$

where X^T is the transpose of matrix X and X^{-1} is the inverse of matrix X. An important assumption for this solution to hold is that the matrix $X^T X$ is non-singular or invertible [8]. This form of the OLS estimates gives more insight into the limitations of OLS, discussed in the next section.

2.2 Properties and Limitations of OLS

The OLS estimators are the best linear unbiased estimators (BEST), under the assumptions of the Gauss-Markhov theorem [8]. This means that the parameter estimates are unbiased and give the least variance out of all unbiased estimators. However, in some situations these assumptions do not hold. Firstly, OLS faces problems such as multicollinearity, which happens when the β_j coefficients are highly correlated with each other. In this case, the OLS estimate is no longer the best estimator. Mathematically, this happens because the matrix $(X^T X)$ from Equation (6) is no longer singular (or ill-conditioned), so that the estimators have very high variance and exhibit numerical problems. Hence, OLS is an unstable solution [9]. Moreover, when the model contains a large number of predictors, the coefficient estimates are variable and cannot be computed [4]. Further reasons why OLS need to be modified are the accuracy of the model and the interpretability of the model. OLS estimates are often difficult to interpret when the model contains many predictors. In this situation, OLS exhibits low bias and a large variance, which influences negatively the prediction accuracy [9].

This raises the necessity of finding some alternatives to the OLS technique. We need a technique to select only the relevant variables to include in the model, reduce the variance in such a way that the model has a good accuracy and is easy to understand. One class of these techniques is known as shrinkage or regularization [7]. These methods shrink the coefficients towards zero, reduce the variance and select which coefficients to bring or not to zero. Two regularization techniques are ridge regression and lasso [5].

3 Ridge Regression

Ridge regression solves some of the shortcomings of linear regression. Ridge regression is an extension of the OLS method with an additional constraint. The OLS estimates are unconstrained, and might exhibit a large magnitude, and therefore large variance. In ridge regression, the coefficients are applied a penalty, so that they are shrunk towards zero, this also having the effect of reducing the variance and hence, the prediction error. Similar to the OLS approach, we choose the ridge coefficients to minimize a penalized residual sum of squares (*RSS*). As opposed to OLS, ridge regression provides biased estimators which have a low variance [4].

3.1 Bias Variance Trade-Off

In order to understand the improvement of ridge regression over OLS, we first look at the relation between bias and variance. The bias represents the extent to which the expected prediction is different from the value we are actually predicting. Mathematically this is equivalent to the squared difference between the true mean and the expected value of the estimate. The variance represents how much the predictions for an individual data point varies around their average when measurements are repeated [1].

Bias and variance strongly affect the prediction error. This is apparent by looking at the Mean Squared Error (MSE), a popular estimation quantity, which is defined as

$$MSE = Bias^2 + Variance. \tag{7}$$

Bias and variance also have an effect on the model complexity. Usually, the issue of model complexity (high number of predictors) is dealt with by dividing the data into a training and a validation set (or test set) and by estimating the coefficients from the training set [1]. When a model contains a large number of parameters, the complexity increases. This has the effect of increasing the variance and decreasing the bias. In the other case, when the model complexity decreases, the variance decreases at the cost of increased bias. We choose the appropriate model complexity such that bias trades off for variance in a way that reduces the error on the test set [3]. The over fitting phenomenon is often a consequence of model complexity. Over fitting means that a model performs good on the training set but poorly on the test set. Therefore, the relation between variance and bias strongly influences over-fitting and under-fitting. Figure 1 depicts how the variance and bias vary as model complexity is modified.

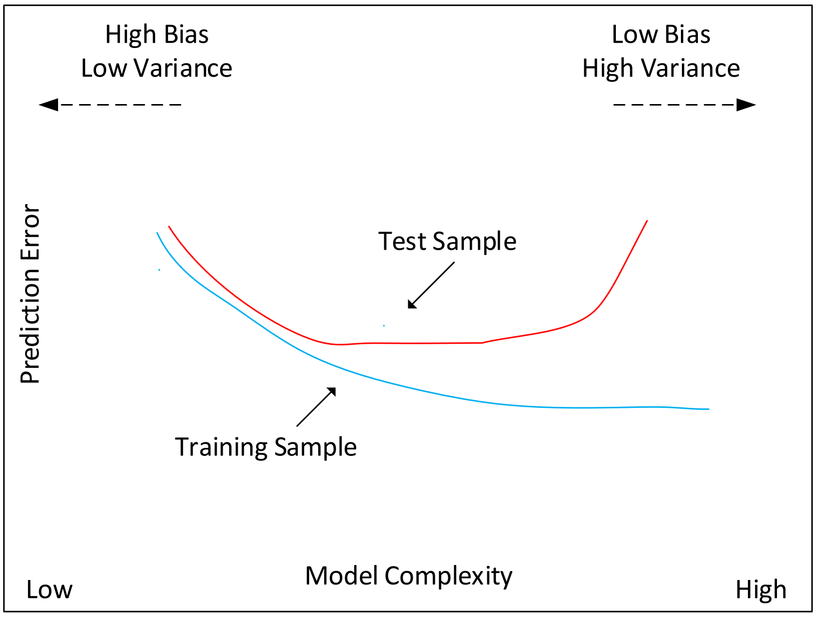

Figure 1: The Bias-Variance Trade-off Achieved with Ridge Regression: Influence on Prediction Error and Model Complexity (adapted from [3]).

When the model complexity is high, we have a large error on the test set, and the predictions exhibit a large variance. If the model complexity is low, under-fitting occurs, resulting in large bias. Typically, the model with the best predictive capability achieves a balance between bias and variance [1]. Ridge regression reaches a trade-off between bias and variance. In the following sections, we show that it produces biased estimates with a large variance and that it works well in situations when there is a large number of predictors. This solves the problem of variability in OLS, which exhibit low bias but high variance [3].

3.2 Mathematical Formulation

This section provides the mathematical formulation of ridge regression. In order to understand its advantages over OLS, we examine three mathematical ways to express the ridge coefficients. The coefficients for the ridge regression are obtained from minimizing the RSS, with an additional constraint given by

$$\hat{\beta}_{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{j=1}^{N} (y_i - (\beta_0 + \sum_{j=1}^{N} \beta_j x_j))^2 + \lambda \sum_{j=1}^{N} \beta_j^2 \right\} \tag{8}$$

$$= \underset{\beta}{\operatorname{argmin}} (RSS + \lambda \sum_{j=1}^{N} \beta_j^2), \tag{9}$$

where the term $\lambda \sum_{j=1}^{N} \beta_j^2$ is known as a "shrinkage penalty", λ is the tuning parameter which we discuss in the following section and $\sum_{j=1}^{N} \beta_j^2$ is the square of the norm of the vector β. This is known as the l_2 norm and is defined as $||\beta||_2 = \sqrt{\sum_{j=1}^{N} \beta_j^2}$. In other words, the ridge coefficients $\hat{\beta}_{\text{ridge}}$ minimize a *penalized RSS*, and because the penalty is given by the l_2 norm, we call it an L_2 penalty [9].

Before finding the parameters $\hat{\beta}_{\text{ridge}}$, we consider two important assumptions for ridge regression. Firstly, the intercept is not penalized. Secondly, the predictors need to be standardized. In contrast with OLS estimates, where multiplying by a constant will scale the coefficients inverse proportionally; the ridge coefficients can change drastically when there is a multiplication with a constant. Hence, for each x_{ij} from the training data, we subtract its mean then divide it by the standard deviation [3]. Therefore, standardizing the inputs yields an intercept and predictors given by

$$\beta_0 = \bar{y} = \sum_{j=1}^{n} \frac{y_i}{n}, \tag{10}$$

$$\overline{x_{\text{ij}}} = x_{\text{ij}} / \sqrt{\frac{1}{n} \sum_{j=1}^{n} (x_{\text{ij}} - \bar{x}_{\text{j}})}, \tag{11}$$

where $\bar{y}$, $\overline{x_{\text{ij}}}$, $\bar{x}_{\text{j}}$ represent the standardized values.

Next we consider ridge regression in matrix algebra representation. We take an input matrix X, which is an n by p matrix with centered inputs, and y a centered n vector. The main convention in ridge regression is that the intercept is left out of the equation. In addition, matrix X is centered, with zero mean and variance unity, and the response y is also centered, with zero mean. After standardizing the inputs, we express the minimization problem in Equation (9) in terms of matrices, as an L_2 penalized problem

$$\hat{\beta}_{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2} ||y - X\beta||^2 + \lambda ||\beta^2||. \tag{12}$$

The ridge coefficients become

$$\hat{\beta}_{\text{ridge}} = (X^T X + \lambda I)^{-1} X^T y, \tag{13}$$

where I is the identity matrix. This shows that, opposed to OLS, ridge regression always provides an unique solution. This is because the quantity $(X^T X + \lambda I)$ is always invertible even in the case when the matrix $X^T X$ non-singular. This argument was the starting point of ridge regression [4].

Another equivalent formulation of ridge regression is obtained by minimizing a constrained version of the *RSS*

$$\hat{\beta}_{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \sum_{j=1}^{N} (y_i - (\beta_0 + \sum_{j=1}^{N} \beta_j x_j))^2, \tag{14}$$

subject to

$$\sum_{j=1}^{N} (\beta_j)^2 < t, \tag{15}$$

where t is a shrinkage factor, $t > 0$. Hence, we obtain the ridge coefficients by minimizing the RSS subject to a constraint given by an L_2 penalty [9].

3.3 Significance of Lambda

In the following, we discuss the importance of the parameter *lambda*, the shrinkage or tuning parameter. The variable λ controls the amount of shrinkage in the ridge estimates and the size of the coefficients. When its value increases and approaches ∞, the $\hat{\beta}_{\text{ridge}}$ coefficients are shrunk towards zero. When its value is 0, we obtain the usual OLS estimates. For every value of λ, a different set of regression coefficients will be generated, as opposed to OLS which provides only one set of coefficients [5].

The value of this parameter plays an important role for the accuracy of the model and it is usually chosen by cross-validation [5]. This performs by taking a grid search of values for λ and computing the corresponding errors for each value of λ. We consider the value of λ for which the error is smaller and then train the model again with this value [3]. The λ parameter is also responsible for the degree of complexity of the model, which has a direct effect on the degree of over-fitting [1]. When λ increases, variance decreases, however, this increases the bias. When λ increases, the opposite behavior happens.

Another concept strongly connected with the shrinkage parameter and model complexity is the effective degrees of freedom. For OLS, the degrees of freedom is equal to the number of free parameters denoted by p. For ridge regression, the degrees of freedom are defined as a function of λ, or more exactly, as the trace of the following expression

$$\begin{aligned} df(\lambda) &= tr[X(X^T X + \lambda I)^{-1} X^T] && (16) \\ &= \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}, && (17) \end{aligned}$$

where d_j represent the singular values of matrix X (discussed in the following section), or simply the eigenvalues of the matrix $X^T X$. This is a decreasing function of λ. When $\lambda = 0$, meaning no penalization, the $df(\lambda) = p$; but when $\lambda \to \infty$, then the $df(\lambda) \to 0$ as a consequence of the parameters being heavily penalized (or constrained). Thus, the more shrinkage is applied, the lower the degrees of freedom [3]. Degrees of freedom are important in calculating some model selection criteria for estimating λ [2].

3.4 Shrinkage

In this section, we investigate further the nature of the shrinkage performed by ridge regression. Ridge regression performs a constant shrinkage on the coefficients, with the shrinkage given by an amount of λ. Ridge regression includes all of the variables of the model [5]. Like mentioned previously, the ridge estimator is a biased estimator of β, as opposed to the OLS estimate which gives an unbiased estimate. An interesting case is when we consider an orthonormal design matrix X (or orthonormal inputs). An orthonormal matrix refers to an orthogonal matrix, with orthogonal columns of length one, considering the columns are expressed as vectors. In this case, the relationship between $\hat{\beta}_{\text{ridge}}$ and $\hat{\beta}_{\text{OLS}}$ becomes

$$\hat{\beta}_{\text{ridge}} = \frac{\hat{\beta}_{\text{OLS}}}{1+\lambda}, \tag{18}$$

which shows that the ridge coefficients are derived from a scaled version of the OLS coefficients [3]. This relationship further illustrates the main characteristic of ridge regression, which is shrinkage. Ridge regression always shrinks the coefficients towards zero, reducing the variance but introducing additional bias.

We now turn to the relationship between ridge regression and principal components analysis (PCA). PCA refers to a method of explaining the variance-covariance structure of linear combinations of variables [9]. We use the Singular Value Decomposition (SVD) of the matrix X, which is an $(N \times p)$ matrix, to gain more insight into how ridge regression performs shrinkage. We express matrix X as

$$X = UDV^T, \tag{19}$$

where

- U is an $(n \times p)$ orthogonal matrix
- V is an $(n \times p)$ orthogonal matrix
- D is a diagonal matrix with dimension $(p \times p)$ and diagonal elements d_{j}, such that $D = \text{diag}(d_{\text{j}})$.

The values $d_1 \geq d_2 \geq \ldots \geq d_p \geq 0$ are the singular values of the matrix X, which means that X becomes a singular matrix if one or a few of the values of d_j are zero. By replacing the expression of X from Equation (19) into Equation (13), and after some mathematical arrangements, the $\hat{\beta}_{\text{ridge}}$ coefficients are given from the following equation

$$X\hat{\beta}_{\text{ridge}} = \sum_{j=1}^{N} u_j \frac{d_j^2}{d_j^2+\lambda} u_j^T y, \tag{20}$$

where u_j are the columns of matrix U. This expression displays the relation between ridge regression and principal component analysis (PCA). SVD serves as a way to express the principal components of the matrix X, and these are in fact the columns of matrix V. The main idea behind PCR is that the largest principal component is the one with the largest variance of the data and the smallest principle component is the one with the smallest variance [3]. Ridge regression calculates the coordinates of y subject to the orthonormal matrix U, and then shrinks them by the factor $\frac{d_j^2}{d_j^2+\lambda}$. As $\lambda \geq 0$, the term $\frac{d_j^2}{d_j^2+\lambda} \leq 1$. This shows that ridge regression shrinks the principal components which correspond to $d_j^2 = \lambda_j$. More exactly, ridge regression shrinks the low-variance directions more and keeps all principle components directions unchanged [9].

4 Lasso

Lasso, or "Least Absolute Shrinkage and Selection Operator", is another regularization method with two additional features to ridge regression. Unlike ridge regression, it shrinks some coefficients exactly to zero. This property is known as sparsity. In addition, lasso shrinks some specific coefficients. Lasso has the property of selecting variables from a large set, property known as variable selection. Therefore, lasso performs regularization and variable selection [7].

4.1 Mathematical Formulation

In this section we discuss several mathematical formulations for the lasso. The lasso problem can be written in the Lagrangian form as

$$\hat{\beta}_{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{j=1}^{N} [y_i - (\hat{\beta}_0 + \sum_{j=1}^{N} \hat{\beta}_j x_j)]^2 + \lambda \sum_{j=1}^{N} |\beta_j| \right\} \tag{21}$$

$$= \underset{\beta}{\operatorname{argmin}} (RSS + \lambda \sum_{j=1}^{N} |\beta_j|), \tag{22}$$

where, as before, λ represents the shrinkage parameter. The term $\sum_{j=1}^{N} |\beta_j|$ is called the shrinkage penalty, and is given in fact by the l_1 norm of the vector β, defined as $||\beta||_1 = \sum |\beta_j|$. Therefore, we call it an L_1 penalty. Thus, the main difference between ridge regression and lasso is that lasso uses an L_1 penalty unlike ridge regression which uses an L_2 penalty. The difference between an L_1 penalty and L_2 penalty is that the L_1 penalty has the effect of shrinking some coefficients exactly to zero [9].

As in the case of ridge regression, the predictors are standardized and the intercept is left out of the model, being estimated as $\beta_0 = \bar{y}$. We express the lasso estimate solution as an L_1 optimization problem

$$\hat{\beta}_{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \sum_{j=1}^{N} [y_i - (\hat{\beta}_0 + \sum_{j=1}^{N} \hat{\beta}_j x_j)]^2 \tag{23}$$

subject to

$$\sum_{j=1}^{N} |\beta_j| < t, \tag{24}$$

where $t > 0$ is a shrinkage factor. Between the parameter λ and t there is a one-to-one correspondence, more exactly t shows the amount of shrinkage that is applied to the parameters. This is in fact a quadratic programming problem, with several algorithms available for solving it [7], like for example the least angle regression (LAR) algorithm.

In order to gain more insight into the properties of the lasso estimates, we look at the matrix algebra formulation. In this case, lasso coefficients are the solutions to an L_1 penalized problem

$$\hat{\beta}_{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} 1/2||y - X\beta||^2 + \lambda ||\beta||_1, \tag{25}$$

which always provides a unique solution that exists when the matrix $X^T X$ has a full rank. As opposed to ridge regression, the coefficients $\hat{\beta}_{lasso}$ have no closed form, because the constraint given by the L_1 penalty is in absolute value, which cannot be differentiated. The solutions for the lasso problem are nonlinear in y_i because the constraint has a non-smooth nature [9].

4.2 Properties of Lambda

Similar to the ridge regression case, parameter λ (or t) controls the amount of shrinkage or regularization applied to the lasso coefficients. We denote $t_0 = \sum_{j=1}^{N} |\beta_{\mathrm{OLS}}|$, as the sum of the absolute value of the OLS estimates. If $t > t_0$, then no shrinkage is performed and the lasso estimates are just the OLS estimates. If t is chosen such as $0 < t < t_0$, then shrinkage towards zero occurs, with some of the estimates being exactly zero. For example, if $t = t_0/2$, shrinkage by an amount of 50 % occurs in the least squared estimates. Thus, if λ is large enough, or equivalently, t is small enough, the coefficient estimates are exactly zero, as opposed to ridge regression where the coefficients shrink but do not reach zero. We refer to this type of solutions as sparse.
In addition, the parameter λ controls the number of variables to be included in the model. Thus, lasso also performs model selection. Similar to ridge regression, the value of λ affects the accuracy of the model [9]. Unlike for ridge regression, the degrees of freedom are not easily defined. The degrees of freedom are often used to estimate the model complexity. For the case of lasso, the degrees of freedom are an unbiased estimate of the number of nonzero coefficients. The degrees of freedom are usually used for model selection criteria (such as AIC, BIC) to provide an optimal lasso fit [11].

4.3 Shrinkage: Soft Thresholding

The lasso performs a type of shrinkage called soft-thresholding. For this, we consider the orthonormal design matrix X of $n \times p$, and assume that $X^T X = I$ (orthogonal case). Like in the case of ridge regression, we express a relation between the lasso estimate and the OLS estimate

$$\hat{\beta}_{\mathrm{lasso}} = \operatorname{sign} \hat{\beta}_{\mathrm{OLS}} (\hat{\beta}_{OLS} - \gamma)_{+}, \tag{26}$$

where γ is a constant determined from the equation $\sum_{j=1}^{N} |\hat{\beta}_{\mathrm{OLS}}| = t$ [9] and we use the $_+$ sign to denote the positive part of Equation (26). This type of estimator is known as a "soft-threshold" estimator. Soft-thresholding means that the coefficients less than γ are shrunk to zero and coefficients larger than γ are shrunk by the amount of γ. This shows that lasso performs a continuous variable selection[9].

5 Comparison between Ridge and Lasso

The previous chapters discussed how ridge and lasso work from a mathematical point of view and how each method shrinks the coefficients. This chapter emphasizes some similarities and differences between the two methods, in order to differentiate the nature of the shrinkage performed in both methods.

5.1 Geometrical Interpretation

We now compare the shrinkage methods by looking at the geometry of ridge and lasso. Figure 2, adapted from [1], displays the estimation problem for both ridge and lasso, when there are only two predictors. The figure displays the constraint regions from Equation (14) respectively Equation (23) and the elliptical contours centered at the *OLS* estimate represent regions where the

RSS is constant. Both regression methods find the point where the elliptical contour intersects the constraint region. Ridge regression has a circle shaped constraint region defined by $\beta_1^2 + \beta_2^2 \leq t$, while lasso has a diamond shaped constraint region given by $|\beta_1| + |\beta_2| \leq t$. In the case of lasso, if the contour intersects the diamond at a corner, then one of the coefficients β_j is equal to 0. In contrast, for ridge regression, there is no intersection between the contour and the constraint at the axis, which shows that the ridge coefficients will not be exactly equal to zero [5]. This illustrates in a graphical manner the sparsity property of the lasso.

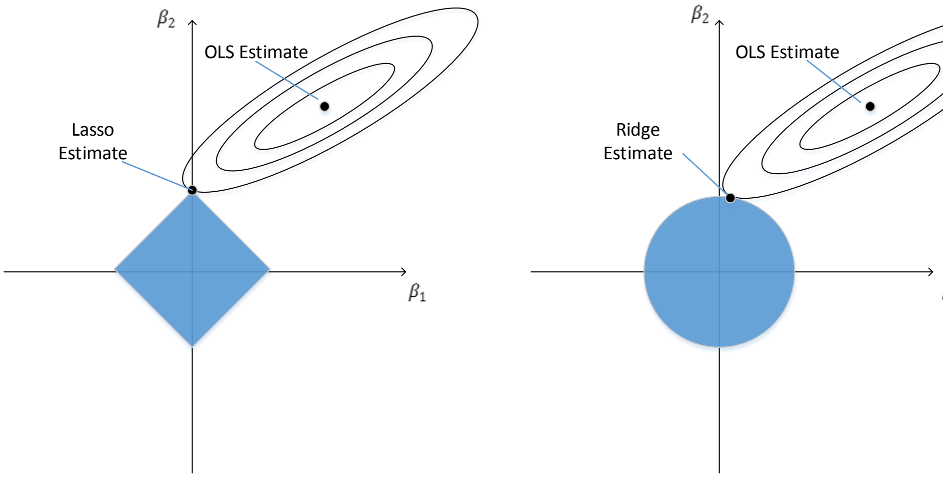

Figure 2: Lasso and Ridge Geometrical Interpretation: Contours of the Errors, represented by elliptical contours and Constraint Functions for ridge ($\beta_1^2 + \beta_2^2 \leq t$) and lasso ($|\beta_1| + |\beta_2| \leq t$).

5.2 Bayesian Interpretation

Another way to look at the sparsity property of the lasso is by defining both the ridge and lasso estimates as Bayesian estimates. We formulate both ridge and lasso as Bayes estimates with different prior distributions [3]. First, we consider a linear regression model

$$y_i = \sum_{j=1}^{N} X_{ij}\beta_j + \varepsilon_i, \tag{27}$$

where ε_i are the errors, which are independent and drawn from a normal distribution, y_i is the dependent variable, X_{ij} represent the indenpendent variables and β_j the regression coefficients. We derive the ridge and lasso estimates from the linear regression model, with an additional assumption, that there is a prior distribution for β_j [3]. For ridge regression, the prior distribution of β_j and of y are

$$P(\beta_{\text{ridge}}) = N(0, \tau), \tag{28}$$

$$P(y) = N(x^T\beta, \sigma^2) \tag{29}$$

where σ^2 is the variance and both σ and τ are known before. The relation between λ and τ is $\lambda = \sigma^2/\tau^2$. If we multiply this distribution by a likelihood function, then the resulting distribution is called the "posterior distribution". We assume that β_{ridge} have a prior Gaussian distribution with zero mean; in addition, λ is a function of the standard deviation. Then, the ridge coefficients are the mean and the mode of the posterior distribution since for the Gaussian distribution the mode and mean coincide [1].

For the lasso case, we consider a double exponential (or Laplacian) prior distribution which has the form

$$p(\beta_{\text{lasso}}) = \frac{1}{2}\tau(\exp(-|\beta_j|/\tau), \tag{30}$$

where $\tau = 1/\lambda$, or, equivalently $p(\beta_{\text{lasso}}) = \lambda/2(\exp(-\lambda|\beta_j|)$ [6]. Hence, the lasso estimates are proportional to the log density of the double exponential distribution. We derive the lasso estimate as the posterior mode with an independent double-exponential prior. However, this is not the posterior mean as in the ridge regression case [7]. The shape of the Laplacian distribution of the lasso explains the sparsity property of lasso. the Laplacian distribution has a peak point, which indicates that some of the coefficients shrink exactly to zero [3], as opposed to the Gaussian distribution which has a bell-shaped distribution.

We conclude this section by a generalized form of both ridge and lasso estimates in the Bayesian interpretation, considering the criterion

$$\widetilde{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j|^q \right\}, \tag{31}$$

The term $|\beta_j|^q$ represents the prior distribution and the quantity $\sum_{j=1}^{p}|\beta_j|^q$ is the contour of the prior distribution of the parameters, also called the L_q norm. The case of $q = 1$ corresponds to the lasso, with the Laplacian distribution and $q = 2$ corresponds to ridge regression [5]. Figure 3 displays the contours of the shrinkage term (or regularization term).

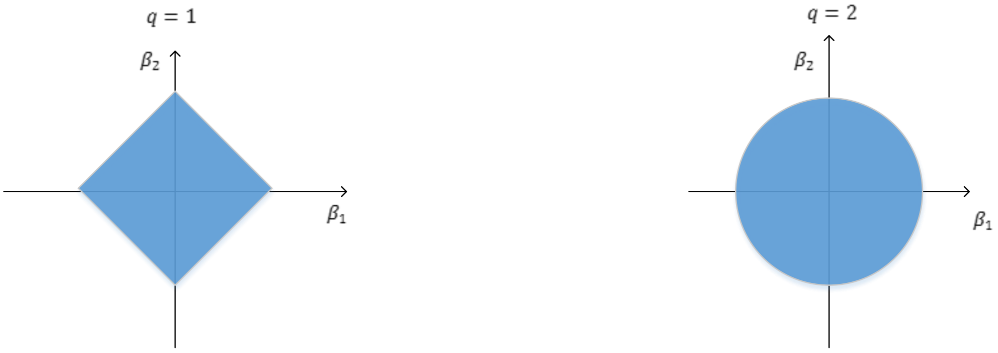

Figure 3: Contours of the Penalization Term for Lasso (left) and Ridge (right) (adapted from [5]).

5.3 Discussion

In this section, we briefly present the advantages and disadvantages of both methods. A comparative overview of the two methods can be observed in Table 1.

Both ridge and lasso shrink the OLS estimates by a certain amount, by penalizing the Residual Sum of Squares (RSS). Lasso measures the shrinkage by $\sum_{j=1}^{N}|\beta_j|$, while ridge by $\sum_{j=1}^{N}(\beta_j)^2$.

Criteria	**Lasso**	**Ridge Regression**
Shrinkage Amount	$\sum_{j=1}^{N} \lvert\beta_j\rvert$	$\sum_{j=1}^{N} (\beta_j)^2$
Shrinkage Type	Proportional	Soft-thresholding
Orthogonal design	Equation (18)	Equation (26)
Number of variables	Controlled by λ	All variables
Performance	Few predictors, high coefficients	Many predictors, coefficients same size
Advantages	More interpretable	Good accuracy
Limitations	Large no of predictors, no group selection	Model not parsimonious

Table 1: Comparison between ridge regression and lasso.

Thus, the two methods use different kinds of penalties applied to the OLS equation. Ridge performs shrinkage in a proportional manner, while lasso applies a type of shrinkage called *soft thresholding*, which shrinks coefficients with a fixed quantity. In case of orthogonal design, ridge and lasso estimates are simple functions of the OLS estimates, given by Equation (18), respectively Equation (26). Figure 4 depicts this relationship graphically. It displays the OLS, ridge and lasso estimates plotted against the OLS estimates and the respective model estimates. We can see that both the ridge and lasso estimates are functions of the OLS estimates.

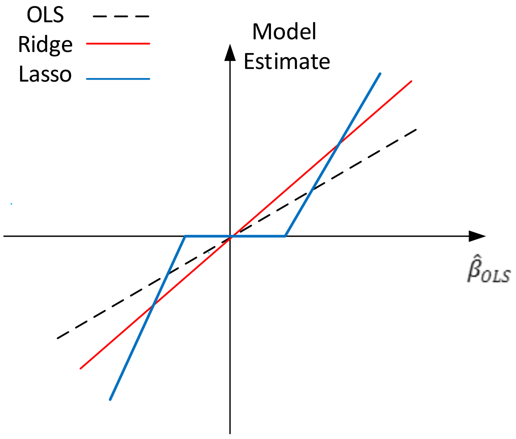

Figure 4: Ridge and Lasso Estimates in the Orthonormal Case (adapted from [5]).

For lasso, parameter λ controls how many variables to include in the model, while ridge includes all the predictors in the model, thus not resulting in a parsimonious model (as small as possible). Both methods outperform OLS because both achieve a reduction in variance at the cost of an increase in bias. Lasso coefficients estimates are more interpretable, as a consequence of the variable selection feature. However, in terms of prediction error (MSE) the two methods are not comparable and we cannot determine which method performs better. According to [5], lasso performs well in the setting when very few of the predictors have high coefficients and the rest very low coefficients. Ridge regression performs well when there are many predictors to explain the output, and each of the coefficients associated with the predictors have approximately the same size. The choice of these methods depends on the particular data set at hand [3]. Nevertheless, the lasso method is more popular due to its variable selection property.

However, lasso exhibits certain limitations. In the case when the number of predictors is much higher than the number of observations, lasso does not perform well. Some of the problems in this situation include the fact that lasso selects at most n variables, and that for the lasso to be defined we need to specify a bound on the L_1 norm. Furthermore, the lasso cannot select a whole group of variables, in the cases when there are correlations between variables in the group [10]. Lastly, in [7] it is demonstrated that when $n > p$ and the predictors are highly correlated, the lasso prediction drops, and ridge outperforms it.

Extensions to improve the lasso method have also been developed recently. One of these techniques is the *elastic net* [10] [3], which combines ridge regression shrinkage and lasso variable selection property and has the additional advantage of grouping variables. Other extensions include *BRIDGE* regression, which introduces a generalization of l_q norms or the garotte [7].

6 Conclusion

In this paper, we explained the shrinkage methods of ridge regression and lasso. We use these methods in business forecasting, when we try to predict future data based on past observations. Shrinkage methods achieve a better prediction accuracy. Both ridge and lasso shrink the value of the coefficients towards zero. Shrinkage solves some of the problems of OLS estimates, such as multicollinearity, and helps reduce problems associated with complex models, by avoiding over-fitting. The methods minimize a penalized residual sum of squares with different penalties. The amount of shrinkage is controlled in both methods by a parameter usually chosen by cross-validation. Ridge regression applies an amount of shrinkage which brings the coefficients towards zero. This reduces the variance, at the cost of increased bias and improves the accuracy of the model. Lasso soft thresholds the coefficients exactly to zero, which yields sparse models. In addition, lasso performs variable selection and provides models that are easier to interpret. The choice of which method to use depends on the data set. However, lasso exhibits low performance, for example when the number of predictors is much higher than the number of observations or when choosing grouped variables. There are several algorithms available for computing the lasso estimates and several extensions to ridge and lasso. Some include elastic net, BRIDGE, or the garotte. We can conclude that ridge and lasso are two powerful tools used for regression, with ridge outperforming OLS and the elastic net outperforming lasso.

A References

[1] C. M. Bishop. *Pattern recognition and machine learning*. New York: Springer, 2006. ISBN: 0387310738.

[2] I. E. Frank and J. H. Friedman. *A Statistical View of Some Chemometrics Regression Tools*. In: Vol. 35, No. 2 (May 1993), p. 109. ISSN: 00401706.

[3] J. Friedman, T. Hastie, and R. Tibshirani. *The elements of statistical learning*. Vol. 1. Springer Series in Statistics New York, 2001. ISBN: 978-0-387-84858-7.

[4] A. E. Hoerl and R. W. Kennard. *Ridge Regression: Biased Estimation for Nonorthogonal Problems*. In: Vol. 42, No. 1 (Feb. 2000), p. 80. ISSN: 00401706.

[5] G. James et al. *An Introduction to Statistical Learning*. Vol. 103. Springer Texts in Statistics. New York, NY: Springer New York, 2013. ISBN: 978-1-4614-7137-0.

[6] T. Park and G. Casella. *The Bayesian Lasso*. In: Vol. 103, No. 482 (June 2008), pp. 681–686. ISSN: 0162-1459, 1537-274X.

[7] R. Tibshirani. *Regression Shrinkage and Selection via the Lasso*. In: *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 58, No. 1 (1996), pp. 267–288. ISSN: 0035-9246.

[8] J. Wooldridge. *Introductory Econometrics: A Modern Approach*. Cengage Learning, Sept. 26, 2012. 910 pp. ISBN: 1111531048.

[9] X. Yan, X. Su, and World Scientific (Firm). *Linear regression analysis theory and computing*. Singapore; Hackensack, N.J.: World Scientific Pub. Co., 2009. ISBN: 9789812834119.

[10] H. Zou and T. Hastie. *Regularization and variable selection via the elastic net*. In: Vol. 67, No. 2 (2005), pp. 301–320.

[11] H. Zou, T. Hastie, and R. Tibshirani. *On the degrees of freedom of the lasso*. In: Vol. 35, No. 5 (Oct. 2007), pp. 2173–2192. ISSN: 0090-5364.

B List of Figures

C List of Tables