Stockholms universitet

A simulation study of model fitting to high dimensional data using penalized logistic regression

Ellinor Krona

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2017:6
Matematisk statistik
Juni 2017

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Bachelor Thesis **2017:6**
http://www.math.su.se

A simulation study of model fitting to high dimensional data using penalized logistic regression

Ellinor Krona*

June 2017

Abstract

Preceding studies show that some common variable selection methods do not conform with high dimensional data. The purpose of this study is to introduce reduction of high dimensional data using penalized logistic regression. This study evaluates three penalization methods; ridge regression, the lasso and the elastic net for model fitting on four simulated examples of high dimensional data sets. For each example 30 data sets were simulated containing 400 predictors and 200 observations. Each example differed in correlation among predictors and relation to the binary response variable. Descriptive statistics and measures of predictive power were used to analyze the methods. The results showed that for high dimensional correlated data the elastic net and ridge regression dominate the lasso regarding the predictive power. There were significant differences (P-value $<$ 0.01) when comparing the predictive power using AUC between the methods in 2 out of 4 examples. In conclusion, the elastic net is notably useful in $p \gg n$ case. In addition, the lasso is not a satisfactory method when p is much larger than n. Ridge regression is proved to have high predictive power but is refrained from shrinking coefficients to be exactly zero.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: ellinor.krona@gmail.com. Supervisor: Jan-Olov Persson.

Acknowledgements

Foremost, I would like to thank Jan-Olov Persson for guidance throughout the writing of this report. I would further like to thank my family for their support.

Contents

1 Introduction

Logistic regression [1] is a popular method to model binary classification problems. In the beginning, its application was mostly seen in biostatistics, however the application has spread to areas such as credit scoring [2] and genetics [16]. Over the years, logistic regression has become one of the most important models for categorical response data.

In statistical theory it is often assumed that the sample size is much larger than the number of predictor variables. Large asymptotic theory is then used to derive procedures and assess model accuracy. However, in the high dimensional setting where the number of variables exceed the number of observations, the large asymptotic theory assumption is violated [7], [5]. With advances in technology, high dimensional data is becoming more frequent. In many applications the response variable is related to a small number of predictor variables among a large number of possible variables. Therefore, variable selection is important to identify the relevant variables in high dimensional data. One attractive method is to use penalized logistic regression [9].

Similar to ordinary maximum likelihood estimation, penalized logistic regression estimates the coefficients by maximizing the log-likelihood function, but with subject to a function that imposes a penalty on the size of the coefficients. The penalty causes the coefficients estimates to be biased, but by decreasing the variance of the coefficient estimates it improves the prediction accuracy of the model [13]. The penalty forces the coefficients to shrink towards zero, that is why penalized logistic regression sometimes is referred to as shrinkage or regularization methods. In this thesis we will cover three methods for penalized logistic regression; ridge regression [17], the least absolute shrinkage and selection operator (*the lasso*) [25] and the elastic net [28].

Ridge regression improves prediction error by shrinking large coefficients to reduce overfitting. In a paper by Le Cessie et al. [19] it is discussed how ridge estimators can be combined with logistic regression to improve the model under certain conditions. The lasso was originally proposed for linear regression models by Tibshirani [25]. In contrast to ridge regression, the lasso tends to reduce overfitting and simultaneously perform selection of predictor variables. Several extensions of the lasso such as the group lasso [20] and the relaxed lasso [21] have been proposed. Later on, a new regularization and variable selection method called the elastic net was proposed by Zou et al. [28]. It was shown that the lasso had some limitations [28] regarding variable selection and the elastic net regularization was proposed to overcome these. Elastic net regularization has been applied to portfolio optimization [24] and genomics [4] amongst others.

The purpose of the study is to introduce reduction of high dimensional data using penalized logistic regression. This study evaluates the three penalization methods; ridge regression, the lasso and the elastic net for model fitting. A simulation study is conducted to do a comparative analysis of the penalization methods on four examples of high dimensional data sets. We want to investigate how these methods can be applied to logistic regression to improve the parameter estimates and diminish the error made by further predictions.

To begin with, we provide the statistical theory in Section 2. It is followed by specific statistical theory that is relevant for the modelling and the comparative analysis of the methods. In Section 3 we introduce the simulated data and apply the theory to compare the shrinkage methods. In Section 4 we provide the most important results, that is later on discussed and analyzed in Section 5.

2 Theory

The theory section will cover the necessary theory to explain the modelling and simulation that is later on conducted in the report. In the beginning of every subsection the main source will be referred to. If nothing is mentioned, the reference will appear throughout the report.

2.1 High dimensional problems, $p \gg n$

Many traditional statistical methods for classification are generally intended for problems with a large sample size and a lower dimension. Here, dimension refers to the number of predictor variables, p. In the recent period of time, the collection of data has changed in fields such as finance, marketing and medicine [14]. It is prevalent for businesses and researchers to have access to large amounts of data associated with each object or individual. Hence, the dimensionality of the data is very high. Such data sets, that contains more predictors than observations, are referred to as high dimensional [18], [14].

In high dimensional data it is probable that the predictors suffer from multicollinearity [1]. Multicollinearity occurs when several variables in a regression model are correlated. If the variables are correlated, any variable in the model can be expressed as a linear combination of all the other variables in the model [7]. Multicollinearity tends to increase the variance of the regression coefficients and the more variance they have, the more difficult it is to interpret the coefficients. As a result, it is difficult to determine which predictors that are related to the response [18].

In general, adding predictors to the model that are associated with the response will improve the fitted model and lead to a lower prediction error. Though, adding predictors that are not associated with the response will lead to an increase in prediction error. Adding such predictors increase the dimensionality and aggravate the risk of overfitting without improving the prediction error [13]. Including a large number of predictors can lead to improved predictive models if they are associated with the response. Otherwise, they will lead to worse results.

2.2 Variable selection and shrinkage methods

Variable selection methods aim to find the best subset of predictors for the final model. Methods such as forward- and backward stepwise selection [13] can not be used for high dimensional data. Specifically, when the number of predictor variables are large or if the predictor variables are highly correlated [9]. Backward selection is limited to be used when $n > p$ and forward selection is not computationally possible if the amount of data is large. Another method, best subset selection, has been shown to be inappropriate when $p > 30$ since the number of all possible subsets for high dimensional data would be exponentially large [25].

Dimension reduction methods such as principal components analysis (PCA) can also be applied to high dimensional data. PCA is a popular approach for reducing the dimension of a data set. It uses an orthogonal transformation to convert possibly correlated variables into a set of linearly uncorrelated variables called principal components [18]. In this study the main scope is to introduce penalized logistic regression methods, but we refer to [13] for further reading about PCA.

2.3 The bias-variance trade off

From now on, we refer to training data as the set of data used to fit the final model. We refer to test data as the set of data that is used to assess the prediction accuracy of the final model. In the following section all theory is referred to [13].

Model selection is the exercise of choosing a statistical model among several aspirant models. To make a comparison of different models there are some things that need to be considered. Firstly, we want to develop a function that can be used for future predictions of the response variable. Secondly, we want to estimate the prediction error in order to assess how good the constructed model is. One way to estimate the prediction error is to average the misclassification error over the training data, called the training error

(R_t),

$$R_t = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i). \tag{1}$$

Equation 1 refers to the number of incorrect fractions determined by whether the observed data y_i differs from the estimated $\hat{y}_i$. The observed and estimated data is a binary response variable with two possible outcomes 0 and 1. Let $P(Y = 1 \mid \boldsymbol{x}) = f(\boldsymbol{x})$. If the estimated function $\hat{f}(\boldsymbol{x}) \geq 0.5$ then $\hat{y}(\boldsymbol{x}) = 1$ and the reverse is true for $\hat{y}(\boldsymbol{x}) = 0$.

The training error (1) is not a good estimate of the test error. Generally, as the model complexity increase the training error converges to zero. If a very complex model is selected, the model typically overfits the training data and predicts poorly on new observations. The expected prediction error or the

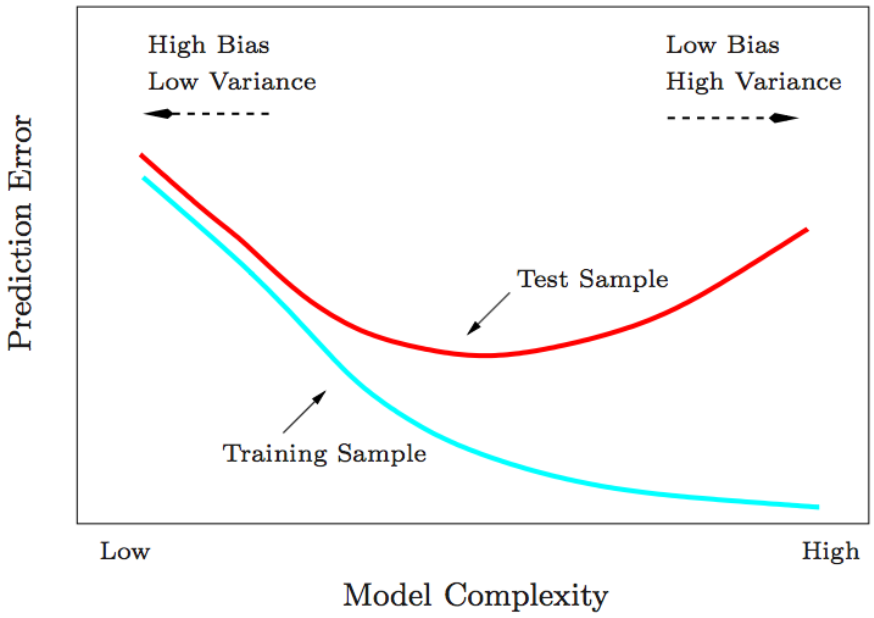

Figure 1: *The bias-variance trade off. The test error and training error is showed as a function of model complexity. As model complexity increase the training error steadily decrease. The test error initially decrease but reaches a minimum because of the bias-variance trade off and then steadily increase. (figure from [13], p.38.) [13]*

test error (R_g) can be expressed as,

$$R_g(x) = \sigma_\epsilon^2 + Var(\hat{f}(x)) + bias^2(\hat{f}(x)), \tag{2}$$

where σ_ϵ^2 is the irreducible error, the variance of the error term ϵ. The second term is the variance and the last term is the squared bias, the measure of how

much the average of the estimate differs from the mean [10]. The test error is calculated by averaging the misclassification error over the observations in the test set.

If a model $\hat{f}(X)$ with high complexity is selected then it is able to follow the relationship between X and Y more closely, resulting in a lower bias but a higher variance. Overfitting will adapt the model too close to the training data and there will be a large test error when predicting on the test data. However, if the model is too simple it will underfit the data and instead have a large bias. The optimal model $\hat{f}(x)$ is chosen such that the variance and bias is minimized simultaneously and gives a minimal test error. Consequently, that is the model that should be selected for future prediction [13].

2.4 Multiple logistic regression

In the following section all theory is referred to [1]. Logistic regression is frequently used to model binary classification problems where the response variable can take one of two outcomes, usually denoted by 0 and 1. In general, the response variable Y is a Bernoulli random variable. The event $Y = 1$ is seen as success and $Y = 0$ as failure. The conditional probability that $P(Y = 1)$ is denoted $\pi(\boldsymbol{x})$ where $\boldsymbol{x}$ is a vector containing the predictors. Binary data frequently result in a non linear relationship between $\boldsymbol{x}$ and $\pi(\boldsymbol{x})$. For such data a multiple logistic regression model is appropriate and the conditional probability is,

$$\pi(\boldsymbol{x}) = \frac{exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}{1 + exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}. \tag{3}$$

The alternative formula, showing the linear relationship by the log odds is,

$$logit[\pi(\boldsymbol{x})] = log\Big[\frac{\pi(\boldsymbol{x})}{1 - \pi(\boldsymbol{x})}\Big] = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p. \tag{4}$$

The log odds transformation is often referred to as the logit. This relates the logit link function to the linear predictor.

Maximum likelihood estimation is used to estimate the regression coefficients of logit models. The likelihood function is maximized to find an estimator that corresponds to the observed data. The likelihood function is given by,

$$L(\boldsymbol{\beta}, y_i) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1 - y_i}. \tag{5}$$

By taking the log of Equation 5, we receive the log-likelihood function,

$$l(\boldsymbol{\beta}, y_i) = \sum_{i=1}^{n} \Big\{ y_i \cdot log(\pi(x_i)) + (1 - y_i) \cdot log(1 - \pi(x_i)) \Big\}. \tag{6}$$

The coefficient estimates $\hat{\boldsymbol{\beta}}$ are obtained by differentiating Equation 6 with respect to $\boldsymbol{\beta}$ and setting the derivatives to zero [1].

2.5 Penalized Logistic Regression

Penalized regression or shrinkage methods are an alternative regression method that involves penalizing the size of the coefficients. Shrinkage methods use a penalty that shrinks the coefficient estimates towards zero. As a result, it improves the prediction accuracy by avoiding overfitting and the resulting model is easier to interpret. Furthermore, it overcomes the problem in high dimensional correlated data [22]. Shrinkage methods are therefore useful to achieve stable and accurate models for high dimensional data.

By incorporating a penalty term in the log-likelihood function (6) the penalized log-likelihood [9] function is obtained,

$$l_p(\beta_0; \boldsymbol{\beta}; \lambda) = -l(\beta_0; \boldsymbol{\beta}) + \lambda P(\boldsymbol{\beta}), \tag{7}$$

where $l(\beta_0; \boldsymbol{\beta})$ is the log-likelihood function seen in Equation 6, $\lambda \geq 0$ is the regularization parameter that adjusts the amount of shrinkage and $P(\boldsymbol{\beta})$ is the penalty function.

The penalized log-likelihood (7) is minimized to find the coefficient estimates. The penalty function is a shrinkage penalty that exhibits a size constraint on the coefficients. When there are a large number of correlated variables in a regression model, their coefficients can be poorly determined and exposed to high variance. A large positive coefficient can be cancelled by a large negative coefficient if the corresponding variables are correlated. The penalty impose a size constraint on the coefficient to allay the cancellation [13].

2.5.1 Regularization parameter

The regularization parameter λ controls the relative effect that the penalty function has on the coefficient estimates. Notice that, the penalty is not applied to the intercept β_0. Penalization of the intercept would make the procedure depend on the origin which is not supported [13]. When $\lambda \to 0$

the penalized log-likelihood (7) converges to the log-likelihood (6). Consequently, the fitted model tends to overfit the data resulting in a model with high variance. When $\lambda \to \infty$ the coefficient estimates approaches zero, the fitted model tend to underfit the data and is too simplistic and may be potentially biased. Thus, the regularization parameter λ directly controls the bias-variance trade off that was earlier described in Section 2.3. In this study, cross-validation (Section 2.6) is used to select λ.

In this report we will consider three penalty functions. In the following three sections we present the penalty functions for ridge regression, the lasso and the elastic net.

2.5.2 Ridge regression

Ridge regression was originally introduced by Hoerl et al. [17]. It was proposed as an alternative to ordinary least squares regression when collinearity was detected among the predictors. Today, it is applied to the logistic regression model as well [19]. Ridge regression solves the following penalized log-likelihood function,

$$l_p(\boldsymbol{\beta}, y_i) = \sum_{i=1}^{n} \left\{ y_i \cdot log(\pi(x_i)) + (1 - y_i) \cdot log(1 - \pi(x)) - \lambda \sum_{j=1}^{p} \beta_j^2 \right\}. \quad (8)$$

Generally, ridge regression includes all predictor variables but with smaller coefficients [13]. In Equation 8 it can be seen that when $\lambda = 0$ the penalized log-likelihood function equals the log-likelihood function without a penalty. The solution to Equation 8 is,

$$\hat{\boldsymbol{\beta}}_{Ridge} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ -l(\boldsymbol{\beta}, y_i) + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}. \quad (9)$$

As a remark, ridge regression does not perform variable selection. It shrinks the coefficient estimates toward zero but is refrained from putting them exactly to zero [17]. As a result, ridge regression is restricted to include all predictor variables in the model. Thus, when p is very large the model is difficult to interpret. Although it does not result in a sparse model, ridge regression has shown to achieve high prediction accuracy when the predictor variables are highly correlated [17]. Since high correlation is frequent in high dimensional data, ridge regression is regularly used [5].

2.5.3 Least Absolute Shrinkage and Selection Operator - The LASSO

The least absolute shrinkage and selection operator (*the lasso*) was originally proposed by Tibshirani [25]. The method maximize the log-likelihood function subject to a constraint on the sum of the absolute values of the regression coefficients. The constraint enables the lasso to perform variable selection by setting some coefficients to be exactly zero. Ultimately, the lasso shrinks some coefficients and sets some of them to zero, which is a combination of best subset selection and ridge regression. The penalized log-likelihood function for the lasso is,

$$l_p(\boldsymbol{\beta}, y_i) = \sum_{i=1}^{n} \left\{ y_i \cdot log(\pi(x_i)) + (1 - y_i) \cdot log(1 - \pi(x)) - \lambda \sum_{j=1}^{p} \left|\beta_j\right| \right\}. \quad (10)$$

As can be seen in Equation 10 there is a similarity to ridge regression. In contrast to ridge regression, the lasso penalty maximizes the log-likelihood subject to the absolute β-value instead of the squared β-value in Equation 8. The solution to Equation 10 is,

$$\hat{\boldsymbol{\beta}}_{Lasso} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ -l(\boldsymbol{\beta}, y_i) + \lambda \sum_{j=1}^{p} \left|\beta_j\right| \right\}. \quad (11)$$

If there is a group of highly correlated predictor variables, the lasso tends to randomly select one predictor variable of that group and neglect the remaining predictors [28]. In addition, the lasso is restricted to select a maximum of n predictor variables in the model [5]. Consequently, when $p \gg n$ it follows that no more than n predictor variables can be included in the model [28]. As a consequence, there could be more than n coefficients that are non-zero, but the predictors are restricted to enter the model [26].

2.5.4 The elastic net

The elastic net was suggested by Zou et al. [28]. It was introduced to compensate for the limitation that the lasso was unable to identify more predictors than the number of observations. Additionally, it promotes a grouping effect [28]. A method exhibits the grouping effect if the coefficients of a group of highly correlated variables are nearly equal.

The elastic net adopts a combination of the penalty terms of ridge regression and the lasso from Equation 8 and 10. The purpose of adding the quadratic part of the penalty (Equation 12) is to remove the limitation on the number

of selected variables [28]. The penalized log-likelihood function for the elastic net is,

$$l_p(\boldsymbol{\beta}, y_i) = \sum_{i=1}^{n} \left\{ y_i \cdot log(\pi(x_i)) + (1-y_i) \cdot log(1-\pi(x)) - \lambda \sum_{j=1}^{p} (1-\alpha)\beta_j^2 + \alpha \left|\beta_j\right| \right\}. \tag{12}$$

The parameter α adjusts the penalty term such that when α is close to zero we obtain the ridge penalty and if α is close to one we obtain the lasso penalty. This linear combination of the lasso and ridge penalty term was suggested as penalty function by Friedman et al. [12]. The solution to Equation 12 is retrieved by using coordinate descent [28]. The solution is given by,

$$\hat{\boldsymbol{\beta}}_{Elastic\ net} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ -l(\boldsymbol{\beta}, y_i) + \lambda \sum_{j=1}^{p} (1-\alpha)\beta_j^2 + \alpha \left|\beta_j\right| \right\} [12]. \tag{13}$$

In contrast to the lasso, the elastic net can select all p predictor variables even though $p \gg n$ [28]. Furthermore, it removes the constraint that was caused by grouped correlated predictors that was mentioned earlier for the lasso in Section 2.5.3. Conclusively, the elastic net encourages a grouping effect that is to eliminate the trivial predictors, but to include all groups of correlated predictors [26].

2.6 Cross-validation

This section provides a general description of cross-validation. In Section 3.3, we describe how cross-validation is used to find the regularization parameter λ and the fitted model.

Cross-validation is validation method for models that is used to estimate the performance of a predictive model. It is a widely used method for estimating prediction error. Moreover, K-fold cross validation is a procedure that set aside one part of the data to fit the model and a different part to test the model [13]. Consider a data set with n observations. For K-fold cross-validation the n observations in the data set is split into K roughly equal-sized subsets. The k-th subset is used as a validation set and the $K-1$ other subsets are combined training sets to fit a model. The model fitted to the training data is then used for prediction on the validation set, the k-th set of the data. This is done for every subset. Figure 2 illustrates the procedure for $K = 5$, where the fourth subset is used as a validation set and the other four subsets are used as combined training sets. By averaging the prediction errors for the K validation sets we obtain the cross-validation estimate of

the prediction error. That is an overall measure of prediction accuracy [13].

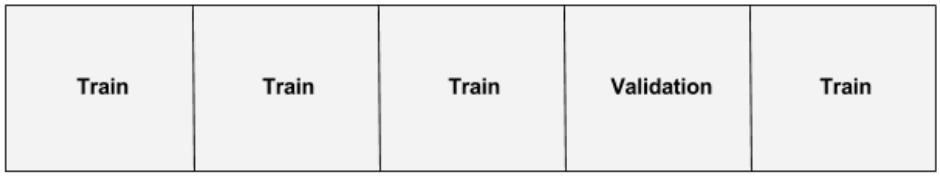

Figure 2: *5-fold cross-validation. The n observations are divided into five subsets. A model is fitted to the validation subset by using the four training subsets. This is done for all five subsets.*

When choosing the optimal K we must consider the bias-variance trade off (Section 2.3). Leave-one-out cross-validation (LOOCV) refers to the case where $K = n$. For LOOCV the bias is low due to subsequent fitting of n subsets. However, LOOCV yields high variance. When we move on to the next observation the previous validation set is included in the new training set. Since this is done for all n subsets, the training sets are overlapping resulting in a higher variance. In this study 10-fold cross validation is used because it has been shown to have some advantages over other choices of K [6]. It is also less computationally intensive than LOOCV for large data sets. For K-fold cross validation, the variance is lower because the number of folds is less, instead the bias is slightly increased.

2.6.1 One-standard-error rule

In the simulation in Section 3.2, the cross-validation error estimates the prediction error at fixed values of the regularization parameter λ. For each λ cross-validation is repeated. We choose λ according to the one standard error rule [25]. We start with the estimate of λ that minimizes the cross-validation error, then we increase λ such that the regularization increase but it remains within one standard error of the minimum,

$$\lambda^* = \min \Big\{ \lambda : CV(\lambda) \geq CV(\hat{\lambda}) + s.d(\hat{\lambda}) \Big\}. \tag{14}$$

Thus, we choose the most regularized model such that the cross-validation error is within one standard error of the minimum. This is motivated by that the λ that achieves the smallest cross-validation error does not yield enough regularization [13].

2.7 Generalized Linear Models

In the simulation study in Section 3 we use generalized linear models. From now on, we will refer to them as GLMs. A GLM is built of three components: a random component, a systematic component and a link function. We will not provide a complete theory about GLMs in this study. For specific information regarding GLMs we refer to [1].

2.7.1 GLM for penalized logistic regression in R

In the simulation study (Section 3) *R* Statistical Software [23] is used. The *glmnet* [11] package for *R* fits a generalized linear model via penalized maximum likelihood. The *glmnet* solves the following problem where the penalized log-likelihood is maximized,

$$\max_{\beta_0,\boldsymbol{\beta}} \frac{1}{N}\sum_{i=1}^{N}[y_i(\beta_0+\boldsymbol{\beta}^T x_i) - log(1+e^{\beta_0+\boldsymbol{\beta}^T x_i})] - \lambda P_\alpha(\boldsymbol{\beta}). \tag{15}$$

The maximization problem consists of the log-likelihood part and the penalization part which is,

$$P_\alpha(\boldsymbol{\beta}) = (1-\alpha)\frac{||\beta_j||_2^2}{2} + \alpha||\beta_j||_1 = \lambda\sum_{j=1}^{p}\left\{(1-\alpha)\frac{\beta_j^2}{2} + \alpha\left|\beta_j\right|\right\}. \tag{16}$$

The regularization parameter λ is chosen over a grid of values, referred to as the regularization path. To find an optimal estimate of λ the regularization path consists of large range of values [12]. For each value of λ, K-fold cross-validation is performed resulting in estimates of the prediction error. In Figure 5 the estimates of misclassification error is shown as a function of $log(\lambda)$. Standard error bars are displayed for each λ, which are the standard errors of the individual misclassification error rates for each of the K parts. The penalty term is controlled by α where $\alpha = 0$ corresponds to ridge regression, $\alpha = 1$ corresponds to the lasso and $0 < \alpha < 1$ corresponds to the elastic net penalty.

2.8 Summarizing the predictive power

In ordinary regression the coefficient of determination R^2 or the multiple correlation R [3] are used as measures of predictive power. For GLMs other measures are proposed [27]. In this section we cover the measures that are used to assess how good a logistic regression model is for prediction.

2.8.1 Contingency table and classification measures

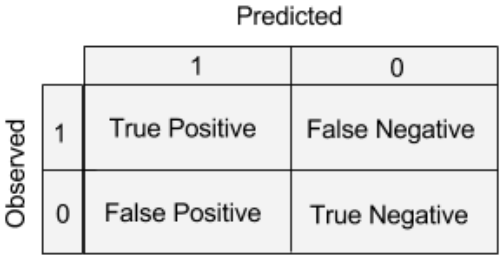

Figure 3: *Contingency table. Table over observed class and predicted class. Categorizing all observations into four classes: true positive, false positive, true negative and false negative. The first row adds to the total positives and the second row adds to the total negatives.*

For each example in Section 3, a contingency table (Figure 3) was produced. The contingency table cross-classifies the predicted value with the observed value. The predicted value is determined by a threshold that, by default, is set to $\pi_0 = 0.5$, such that if $\pi < 0.5$ then $\hat{y}_i = 0$ and if $\pi < 0.5$ then $\hat{y}_i = 1$. Given that we have two classes $Y = 1$ and $Y = 0$, there are four possible outcomes. If the observed value is 1 and it is classified as 1, then it is counted as true positive. If it would be classified as negative it is counted as false negative. True negative and false positive is defined analogously [1]. The true positive rate (TP) of an estimated classifier is,

$$TP = \frac{\#\ true\ positives}{\#\ total\ positives}. \tag{17}$$

In the same way the false positive rate (FP) is,

$$FP = \frac{\#\ false\ positives}{\#\ total\ negatives}. \tag{18}$$

The misclassification error rate (ME) denotes the fraction of incorrect classifications over all observations and is,

$$ME = \frac{\#\ observations\ incorrectly\ classified}{\#\ total\ observations}. \tag{19}$$

2.8.2 Receiver Operation Characteristic

The receiver operation characteristic (ROC) [27] is a description of classification accuracy. The ROC curve plots the TP on the y-axis versus the

FP on the x-axis. Each point in the plot reflects a pair of (FP, TP) for a given threshold. The ROC curve shows how well the classifier distinguishes the two classes for different thresholds, hence it summarizes the predictive power for all possible thresholds.

The optimal result is a concave shaped ROC curve toward the upper left corner. Such a curve implies a high true positive rate and a low false positive rate. On the contrary, a straight line $y = x$ through the origin $(0, 0)$ represents the strategy of a random guess. Thus, if the classifier randomly guess half the time it is expected to get half the positives and half the negatives correct. Such a line is not informative since it reveals no association [1].

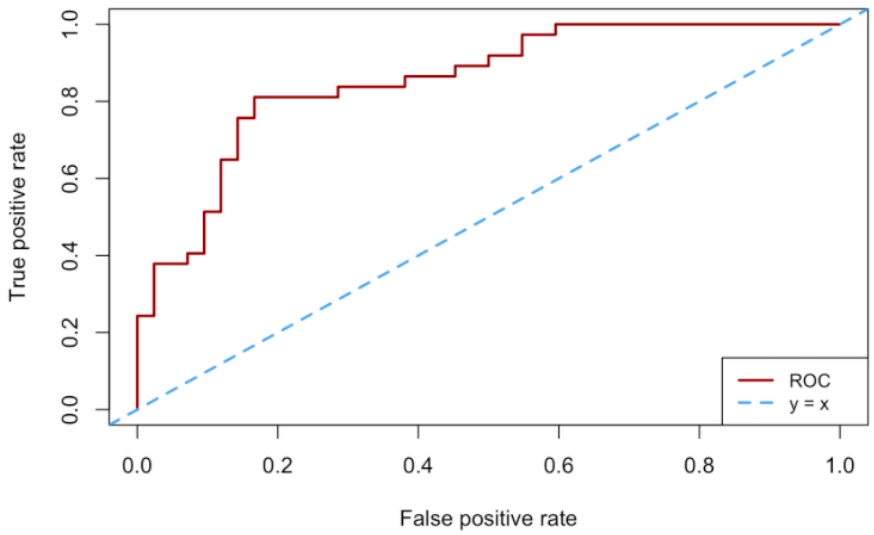

Figure 4: *Plot of a ROC curve and a straight line $y = x$. The true positive rate is plotted against the false positive rate. The ROC curve is concave shaped toward the upper left corner, indicating a better classifier than a random guess.*

The plot in Figure 4 depicts which classifier is the best by noting which one is most skewed toward the upper left corner. However, it is not always easy to determine which classifier is optimal by examining a plot. As a complement, we may compute the area under the ROC curve (AUC) that is another measure of predictive power to compare classifiers.

The AUC is a measure of discrimination and takes values between 0 and 1. The AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive observation higher than a randomly chosen negative observation. An AUC value toward 1 suggests better discrimination. The area under the straight line through the origin equals 0.5. For instance, $AUC \geq 0.9$ implies excellent discrimination and represents a

good classifier [1].

3 Modelling

This section begins with a description of the simulated data set. Then we will present how the simulation is conducted. The theory from Section 2 was applied to perform model selection and validation. We investigated three methods for penalized logistic regression; ridge regression, the lasso and the elastic net. The statistical analysis was implemented using R Statistical Software [23].

The simulated data consisted of four independent high dimensional data sets. Each data set was divided into a training set and a test set. The three methods were used to fit a corresponding model to each of the training sets. The fitted models were used to make predictions for each of the corresponding test sets. Finally, we computed the AUC, the misclassification error and extracted the number of non-zero $\hat{\beta}$-coefficients. The procedure was repeated 30 times per example.

3.1 Simulated data

The purpose of the simulation study was to investigate if there was a difference in the predictive power between the three regularization methods; ridge regression, the lasso and the elastic net, when they were applied to high dimensional data. Furthermore, we considered the interpretability of the model, hence how many variables were selected to be included in the final model. Each method was evaluated using four simulated examples (Section 3.2) of high dimensional data sets.

The simulation study was inspired by the paper by Tibshirani where the lasso was introduced [25]. However, adjustments were made to the simulated data sets. Firstly, we increased the number of predictors such that $p \gg n$ and the data qualified as high dimensional. Specifically, we simulated $p = 400$ and $n = 200$. Secondly, all predictor variables $\boldsymbol{X}$ were continuous multivariate normal distributed except for the binary response variable $\boldsymbol{Y}$. A multiple group of predictors with varying strength of correlation were simulated for each data set.

The predictors were generated by sampling from a multivariate normal distribution with the following probability density function,

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = \Big(\frac{1}{2\pi}\Big)^{n/2} \cdot \frac{1}{\sqrt{det(\boldsymbol{\Sigma})}} \cdot exp\Bigg\{ -\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \Bigg\} \qquad (20)$$

where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ is the covariance matrix. For all $\mathbf{x}$ we set $\boldsymbol{\mu} = 0$ and $Var(\boldsymbol{x}) = 1$. Thus, $\boldsymbol{\Sigma}$ equaled the correlation matrix of $\mathbf{x}$, since by definition the correlation is

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j} \tag{21}$$

and we have set $\sigma_i = \sigma_j = 1$ for all $i, j = 1, 2, ..., 400$. The correlation matrices for Example 1-4 are defined later in Section 3.2 but the general form of a correlation matrix is,

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & \rho_{12} & ... & \rho_{1p} \\ \rho_{21} & 1 & ... & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}. \tag{22}$$

Each predictor variable was assigned a predetermined β-value. Consequently, we obtained a $\boldsymbol{\beta}$ vector that consisted of the corresponding β-values. The response variable were simulated by running the simulated data through the inverse logit function,

$$\pi(\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{X}^T \boldsymbol{\beta}}}. \tag{23}$$

As a result, we retrieved a value of the conditional probability, π. Given the threshold value $\pi_0 = 0.5$ the observed value was categorized into one of two classes; $Y = 1$ if $\pi > 0.5$ and $Y = 0$ if $\pi \leq 0.5$.

Consequently, we obtained a vector $\boldsymbol{Y}$ and a matrix $\boldsymbol{X}$ consisting of 200 observations of the binary response variable and the predictor variables respectively. The simulated data set was divided into a training set $n = 120$ and a test set $n = 80$.

3.2 Simulation

We considered four examples of high dimensional data sets. In this section detailed information about the four examples is provided.

1. In Example 1 we set the pairwise correlation between X_i and X_j predictors to $\text{Corr}(i, j) = 0.5^{\,|i-j|}$. We assigned the first 49 β-coefficients a specified vector that consisted of random values within the range $[2, 5]$. The remaining coefficients were set to 0.

2. In Example 2 we set the pairwise correlation between X_i and X_j predictors to $\text{Corr}(i, j) = 0.5^{\,|i-j|}$. We set all coefficients to be $\beta = 0.85$.

3. In Example 3 the pairwise correlation between X_i and X_j predictors were $\text{Corr}(i,j) = 0.9^{\,|i-j|}$. The coefficients were split into 8 groups, where the coefficients were set to pairwise be 0 and 2,

$$\boldsymbol{\beta} = (\underbrace{2,2,...,2}_{50}\underbrace{0,0,...,0}_{50}\underbrace{2,2,...,2}_{50}\underbrace{0,0,...,0}_{50}...\underbrace{2,2,...,2}_{50}\underbrace{0,0,...,0}_{50}). \quad (24)$$

4. In Example 4 the pairwise correlation between the first 200 predictors, X_i and X_j $(1 < i, j \leq 200)$, were set to $\text{Corr}(i,j) = 0.5^{\,|i-j|}$ and the pairwise correlation for the remaining predictors were set to 0. We set the first 200 coefficients to $\beta = 3$ and the remaining coefficients to 0,

$$\boldsymbol{\beta} = (\underbrace{3,3,...,3}_{200}\underbrace{0,0,...,0}_{200}). \quad (25)$$

3.3 Model fitting and selection

Each one of Example 1-4 was considered separately. Ridge regression, the lasso and the elastic net were fitted to the same data set simultaneously. We used GLMs (Section 2.7.1) to fit the models to the data set.

The regularization parameter λ was found by 10-fold cross-validation. The regularization path for λ for ridge regression and the lasso was defined as 100 values in the range $(10^{-2}, 10^2)$. The regularization parameters (α, λ) for the elastic net were determined by searching through a grid of values that consisted of all possible combinations of λ and α. We let α take 10 values between $(0.05, 0.9)$ and the regularization path for λ consisted of 20 values between $(10^{-2}, 10^2)$. By default, $\alpha = 0$ for ridge regression and $\alpha = 1$ for the lasso (Equation 16).

3.3.1 Model selection algorithm

For λ_j, $j = 1, 2, ..., l$:

1. The training set was randomly divided into 10 roughly equal-sized subsets.

2. One of the three variable selection methods ridge regression, the lasso or the elastic net was chosen to construct a model for prediction of y_i.

   For each $i = 1, 2, ..., 10$ subset samples:

   (a) The $K-1$ subsets were used to fit a model with the chosen method.

   (b) The model was then used to predict y_i for the K-th subset.

The procedure was made for each of the K subsets.

3. Predictions of y_i was made for every subset. The prediction error was calculated as the average cross-validation error over all subsets,

$$CV = \frac{1}{10}\sum_{i=1}^{10} I(y_i \neq \hat{y}_i) \tag{26}$$

where y_i and $\hat{y}_i$ is the binary observed and predicted response value respectively.

4. We choose λ^* according to the one-standard-error rule such that

$$\lambda^* = \min\left\{\lambda : CV(\lambda) \geq CV(\hat{\lambda}) + s.d(\hat{\lambda})\right\}. \tag{27}$$

The fitted model for λ^* was used for prediction on the test set.

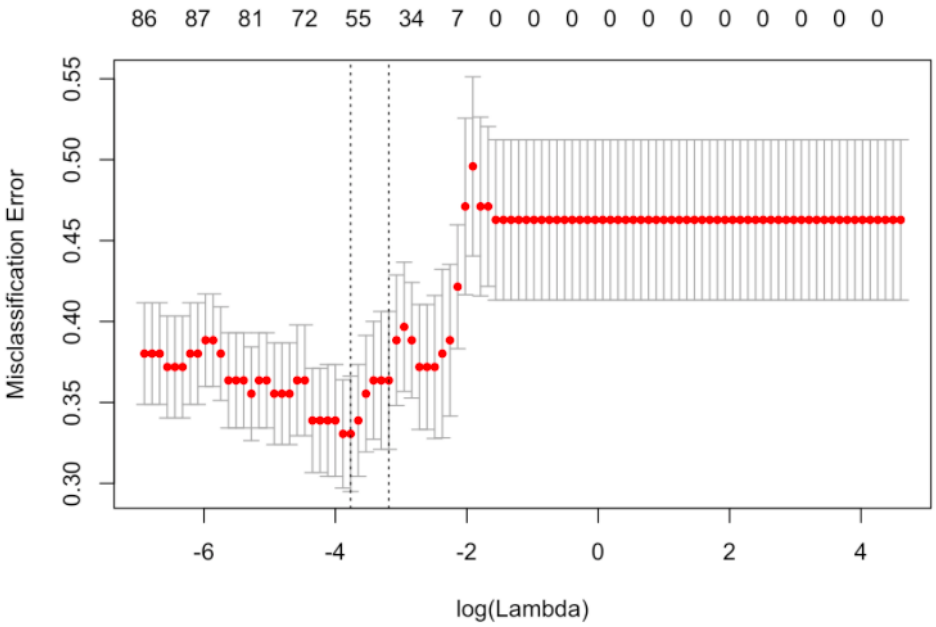

Figure 5: *Cross validation plot for the lasso for one simulation. The cross-validated ME and standard deviations for each value of log(λ). As log(λ) increase, the number of non-zero coefficients decrease as indicated by the axis above the plot. The vertical lines show the locations of the minimum λ and the λ* according to the one-standard-error rule.*

The model selection algorithm was repeated 30 times per example. As a result, we obtained 30 fitted models for ridge regression, the lasso and elastic net respectively. Given the fitted models, we made predictions on the

corresponding test sets. The ME, the AUC and the number of non-zero $\hat{\beta}$-coefficients were calculated and averaged. In order to verify if the differences in AUC and ME between the methods were significant, a non-parametric Friedman's test [8] was conducted. As a complement to the Friedman's test, we performed pairwise comparisons of the methods using Conover's post-hoc test [8]. The Conover's post-hoc test determined which differences were significant.

4 Results

The simulation of Example 1-4 was repeated 30 times. For every simulation we calculated AUC, ME and their standard deviations ($s.d.$). In addition, the average number of selected variables by the lasso and the elastic net was calculated. The results are summarized in Table 1. The results from the Friedman's test and Conover's post-hoc test showed that the pairwise comparisons of AUC were significant ($P\text{-}value < 0.01$) between ridge regression, the lasso and the elastic net in Example 2-4. The difference between ridge regression and the lasso was not significant in Example 1. Moreover, the pairwise comparisons of ME were not significant between the elastic net and ridge regression in Example 3 or between ridge regression and the lasso in Example 1. However, the other pairwise comparisons of ME were significant ($P\text{-}value < 0.01$).

In Example 1, a small subset of predictors were assigned non-zero β-coefficients. On average, the lasso and the elastic net selected 31 and 109 variables respectively. In Table 1 we see that the elastic net had the highest AUC and lowest ME. The Friedman's test showed that the difference between the elastic net and the other methods was significant. Furthermore, the Conover's post-hoc test showed that there was not a significant difference between ridge regression and the lasso.

In Example 2 the predictors were assigned coefficients of $\beta = 0.85$ with relatively high correlation among predictors. As demonstrated in Table 1, ridge regression improved over the other methods considering AUC and ME. As mentioned in Section 2.5.2, ridge regression tends to perform well under the circumstances in Example 2. Moreover, the average number of coefficients for the lasso and the elastic net was 42 and 252 respectively. Due to the one-standard-error rule (Section 2.6.1) the lasso chose a highly regularized model. In this setting, the elastic net identified a larger number of coefficients that were correlated and non-zero. The lasso, on the other hand, resulted in a sparse final model but identified less of the non-zero coefficients. Instead, the chosen model resulted in a high misclassification error (Table 1).

	AUC		ME		$Coefficients$
	Ave.	*sd.*	*Ave.*	*sd.*	$Nr.\ of\ \hat{\beta} \neq 0$
Example 1					
Ridge	0.8275	0.0441	0.2744	0.0466	400
Lasso	0.8242	0.0616	0.2602	0.0633	31
Elastic Net	0.8492	0.0544	0.2430	0.0599	109
Example 2					
Ridge	0.8342	0.0445	0.2601	0.0514	400
Lasso	0.6554	0.0710	0.3862	0.0504	42
Elastic Net	0.8029	0.0526	0.2776	0.0456	252
Example 3					
Ridge	0.9635	0.0171	0.1186	0.0324	400
Lasso	0.9201	0.0292	0.1666	0.0472	35
Elastic Net	0.9586	0.0179	0.1174	0.0318	193
Example 4					
Ridge	0.8433	0.0679	0.2552	0.0664	400
Lasso	0.7315	0.0724	0.3400	0.0662	42
Elastic Net	0.8138	0.0739	0.2658	0.0690	190

Table 1: *Average of AUC, ME-values and number of non-zero $\hat{\beta}$-coefficients for ridge regression, the lasso and the elastic net. The simulation was repeated 30 times for each example and the corresponding values of ME, AUC and number of non-zero $\hat{\beta}$-coefficients were averaged.*

In Example 3 the predictors were divided into 8 groups and pairwise assigned coefficients of 0 and 2. Comparable to Example 1, ridge regression outperformed the lasso and the elastic net in view of the AUC. Since the elastic net and ridge regression performed considerably similar, they seem to perform equally as good in this setting. As discussed earlier (Section 2.5.2), ridge regression included all predictors in the final model and resulted in a less interpretable model. However, the elastic net identified on average 193 non-zero coefficients. Supposedly, the elastic net adopted the grouping effect and correctly identified almost all non-zero coefficients simultaneously as it achieved high prediction accuracy.

In Example 4 the predictors were divided into two groups of equal size that were assigned with $\boldsymbol{\beta} = 3$ and $\boldsymbol{\beta} = 0$ respectively. The first 200 predictors were correlated while the remaining 200 predictors were uncorrelated, hence they were independent of the outcome. As seen in Table 1, ridge regression achieved the highest AUC while the elastic net attained the lowest ME. In addition, the elastic net succeeded to identify approximately all non-zero coefficients as a result of the grouping effect.

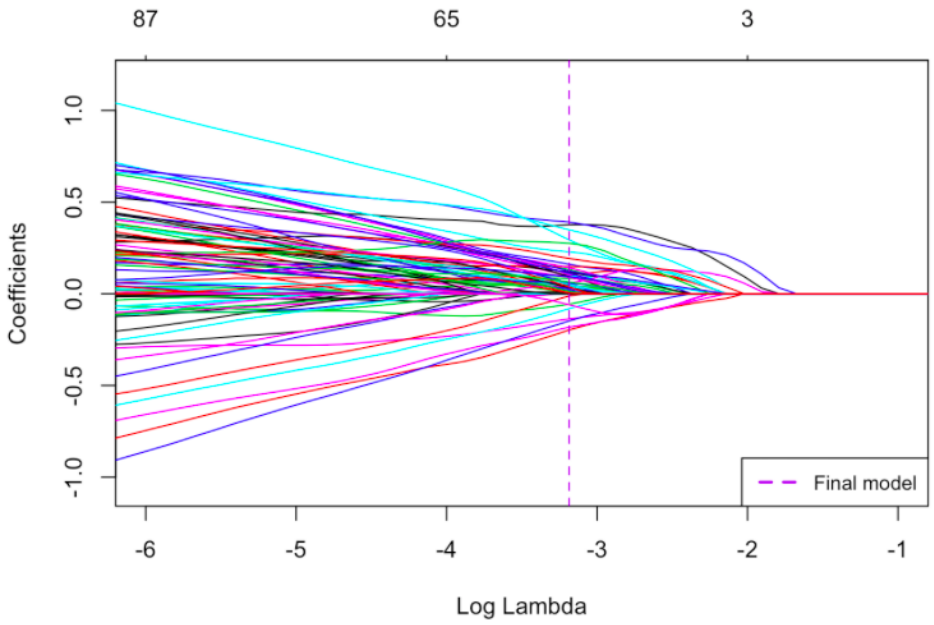

Figure 6: *Coefficient path plot for the lasso for one simulation. Each colored line represents the value taken by a different coefficient in the model. As λ decreases the coefficient size increases as indicated by each line (from right to left). The axis above the plot indicates the number of predictors for different values of λ.*

The number of non-zero coefficients for the lasso and the elastic net through-

out Example 1-4 was determined by the penalty term. Each model generated several coefficient paths for every coefficient. In Section 2.5.1 we discussed how the penalty is imposed on the coefficient estimates. Figure 6 emphasizes the coefficient path for the lasso for one of the simulations in Example 1. The vertical purple line marks the final model corresponding to the optimal value of λ found by 10-fold cross-validation.

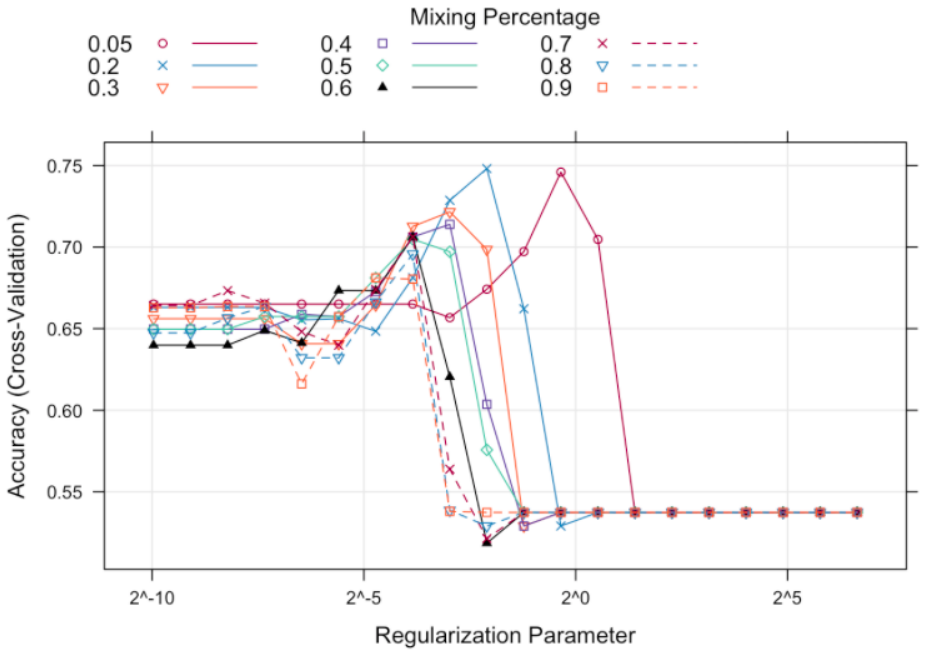

Figure 7: *Cross-validation accuracy versus λ for the elastic net for different mixing percentage. Accuracy refers to the fraction of correctly classified predictions. Mixing percentage refers to the value of α, while regularization parameter refers to λ.*

In Figure 6 each colored line represents the profile of a coefficient in the model. The values of the regularization parameter λ is the weight put on the penalty term. When $log(\lambda) = -2$ all coefficients were essentially zero. As λ approached zero, the coefficients increased from zero and the model approached the ordinary maximization function (Section 2.5). More predictors entered the model when we relaxed λ. When a predictor entered the model it affected other coefficient paths if they were correlated. As a result, some coefficients were reduced to zero as other variables entered the model. The size of a coefficient reduced to zero if other variables apprehended the effect as they entered the model. As a result from the bias-variance trade-off, the optimal λ produced a model with lower complexity. Consequently, the variables that entered the model early were the most predictive and the variables that entered the model later were less important.

As a remark, we observed that the regularization parameter α that controlled the weight of the absolute and quadratic term for elastic net generally was close to zero. In Figure 7 the prediction accuracy for one simulation is plotted against λ for different mixing percentage of α. As demonstrated by the blue line, $\alpha = 0.2$ achieved the highest prediction accuracy. On average, the mixing percentage was less than 0.5 in approximately 90 % of the simulations. Supposedly, the elastic net approached ridge regression by putting more weight to the ridge penalty than to the lasso penalty.

4.1 Summary

The results showed that the three methods performed well in the sense that $AUC \geq 0.5$ in Example 1-4. As discussed in Section 4, the simulations in Example 1 confirmed that the lasso accomplished to quickly identify a small number of important predictors. The results coincided with the conclusions drawn by Tibshirani [25]. Furthermore, we observed that despite the fact that ridge regression tended to spread the coefficient shrinkage over a larger number of coefficients, it achieved high predictive power throughout Example 1-4. Especially the results in Example 3 demonstrated the capacity of ridge regression. We identified that when the number of predictors were very large and a larger fraction of them should be included in the model, ridge regression dominated the lasso and the elastic net. Consequently, it confirmed that ridge regression is a satisfactory method for prediction on correlated data sets [17]. The results from Example 2 determined that the lasso is outperformed by the elastic net. Furthermore, we observed that the elastic net benefits from the adaptability to put a larger weight to the quadratic penalty, while it simultaneously shrinks some coefficients to zero by the absolute penalty.

Moreover, we observed that ridge regression and the elastic net generally improved over the lasso. Specifically, ridge regression dominated the lasso in correlated samples. We asserted that the elastic net approximately identified all non-zero coefficients in the simulations. Generally, the elastic net produced a final model that included all the important predictors. In Example 4 the elastic net performed grouped selection and showed to be a better variable selection method than the lasso. Even though ridge regression did not incorporate variable selection it achieved high prediction accuracy throughout Example 1-4. Therefore, we observed that if the interpretability was not fundamental, ridge regression managed to accomplish high predictive power. Ultimately, the elastic net had the advantage of incorporating variable selection. Consequently, its final model was more interpretable than that of ridge regression.

5 Discussion

In Section 4 we observed that the lasso is outperformed by ridge regression and the elastic net. The study showed that ridge regression is a satisfactory method when the final model should include a larger number of coefficients. Throughout Example 1-4, ridge regression resulted in a final model with high predictive power. However, we showed that the elastic net had high predictive power and produced a relatively sparse, more interpretable model. Moreover, if prediction accuracy is the solemn purpose, ridge regression proved to be one possible good solution. However, variable selection is often of great importance. It is often essential to determine which variables that have significant impact on the response variable. The results showed that the lasso, in general, was not a satisfactory method for high correlated data. The lasso identified a small number of predictors resulting in a sparse model, but it did not achieve high predictive power when it was compared to ridge regression or the elastic net. We observed that ridge regression and the elastic net were better justified approaches for high dimensional correlated data. Instead, the choice of method should be supported by whether variable selection is of importance or not.

In the simulation, we assumed that the partitioning of training and test data yields similar data sets. Since the partitioning of training and test data is random, we assume that it could result in some dissimilar partitions. Furthermore, the partitioning of folds for 10-fold cross-validation is performed randomly. Consequently, we expected fluctuations in the results. As demonstrated by Figure 6, the methods add the predictors individually to the model. The correlated predictors that enter the model later influence the previously included predictors. Thus, the collection of correlated predictors in the final model could be affected by the sampling. Therefore, the simulation was repeated multiple times. Moreover, it would be ideal to increase the number of simulations to be larger than 50. [25] However, it is shown in Table 1 that averaging the results over 30 simulations does not result in large standard deviations. In addition, the Friedman's test and Conover's post-hoc test verified that a principal part of the differences in AUC and ME between methods were significant.

Furthermore, we note that determining the optimal value for the regularization parameter is one of the most relevant problems of penalized regression and can be problematic or result in heavy computation for higher dimensional problems. However, in this study cross-validation was not too computationally intensive and proved acceptable to determine the value of the regularization parameters.

As a remark, we observed that the dimension of the correlation matrix increased as the number of predictors increased. As a consequence, we noted

that the non-degenerate matrix assumption for multivariate normal distribution was almost violated, since the determinant goes to zero [15]. The correlation matrix could therefore cause problems in larger dimensions. For further simulation of higher dimensional problems sampling from a multivariate normal distribution, we advise that this should be taken in consideration.

Before concluding this discussion, we declare that several model selection and validation procedures exist. Ultimately, there is no procedure that outperforms all the others. Generally, different procedures conquer in different situations. In this study, we observed three penalization methods that were used for model fitting to high dimensional data. We showed that they were adequate methods that can be appropriately applied to high dimensional data. Conclusively, they could function as a complement to logistic regression as well as stepwise regression and best subset selection amongst others.

6 Conclusion

In this study we introduced three penalized logistic regression methods; ridge regression, the lasso and the elastic net. We illustrated how the methods could be implemented when analyzing high dimensional data. Emphasis was put not solemnly on the predictive performance of the methods, but also on the removal of predictors that were uncorrelated with the response. It seems that in high dimensional data regularization is an adequate method to achieve good prediction performance. Since ridge regression, the lasso and the elastic net can be applied to much larger data sets than other variable selection methods, they provide a solution for high dimensional classification problems. In conclusion, we showed that the elastic net is notably useful in the $p \gg n$ case. In addition, the lasso is not a satisfying method when p is much larger than n. Moreover, ridge regression is proved to have high predictive power but is refrained from shrinking coefficients to be exactly zero. As a suggestion for future research, this study could be extended to investigate additional penalty functions such as the relaxed lasso [21] or the adaptive elastic net [4]. Furthermore, an amplification of this study could be to include other dimension reduction approaches such as PCA.

References

[1] Agresti, A. (2002): *Categorical Data Analysis.* Second edn. Hoboken, New Jersey: John Wiley & Sons. E-book.

[2] Alburquerque, P.H.M, Medina, F. A.S. & da Silva, A.R. (2017): *Geographically Weighted Logistic Regression Applied to Credit Scoring Models.* Revista Contabilidade & Financas - USP 28(73), 93-112.

[3] Alm, S. E., & Britton, T. (2008): *Stokastik: sannolikhetsteori och statistikteori med tillaempningar.* Liber.

[4] Algamal, Z. Y. & Lee, M. H. (2015): *Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification.* Computers in biology and medicine, 67, 136-145.

[5] Algamal, Z. Y., & Lee, M. H. (2015): *Applying Penalized Binary Logistic Regression with Correlation Based Elastic Net for Variables Selection.* Journal of Modern Applied Statistical Methods. 14(1), 168-179.

[6] Breiman, L. & Spector, P. (1992): *Submodel Selection and Evaluation in Regression. The X-Random Case.* International Statistical Review / Revue Internationale de Statistique. 3, 291-319.

[7] Bühlmann, P. & Van Der Geer, S. (2011): *Statistics for high dimensional data: methods, theory and applications.* Springer Science Business Media. Berlin, Heidelberg: Springer-Verlag. E-book-

[8] Conover, W. J. (1999): *Practical nonparametric statistics.* 3. ed. New York: John Wiley. E-book.

[9] Fokianos, K. (2008): *Comparing two samples by penalized logistic regression.* Electronic Journal of Statistics, 2, 564-580.

[10] Friedman, J. H. (1997): *On Bias, Variance, 0/1-Loss, and the Curse-of-Dimensionality.* Data mining and knowledge discovery, 1(1), 55-77.

[11] Friedman, J., Hastie, Simon, N., T. & Tibshirani, R. (2016): *Glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models.* In: R Package.

[12] Friedman, J. H., Hastie, T. & Tibshirani, R. (2010): *Regularization Paths for Generalized Linear Models via Coordinate Descent.* Journal of statistical software, 33(1), 1. University of California, Los Angeles.

[13] Friedman, J., Hastie, T. & Tibshirani, R. 2008. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction.* Second edn. Springer Series in Statistics. New York: Springer-Verlag. E-book.

[14] Giraud, C. (2015): *Introduction to high dimensional statistics.* Monographs on Statistics Applied Probability, 139. Boca Raton: Chapman and Hall/CRC, EBSCOhost. E-book.

[15] Gut, A. (2009): *An intermediate course in probability.* Springer.

[16] Henshall, J. M. & Goddard, M. E. (1999): *Multiple-trait mapping of quantitative trait loci after selective genotyping using logistic regression.* Genetics, 151(2), 885-894. United States, WAVERLY PRESS INC.

[17] Hoerl, A. E. & Kennard, R. W. (1970): *Ridge regression: Biased estimation for nonorthogonal problems.* Technometrics, 12(1), 55-67.

[18] James, G., Witten, D. & Hastie, T. (2014): *An Introduction to Statistical Learning: With Applications in R.* Sixth edn. New York: Springer-Verlag. E-book.

[19] Le Cessie, S. & Van Houwelingen, J. C. (1992): *Ridge Estimators in Logistic Regression.* Journal of the Royal Statistical Society. Series C (Applied Statistics), 41(1), 191-202.

[20] Meier, L., Van De Geer, S. & Bühlmann, P. (2008): *The Group Lasso for Logistic Regression.* Journal of the Royal Statistical Society. Series B (Statistical Methodology), 70(1), 53-71.

[21] Meinshausen, N. (2007): *Relaxed lasso.* Computational Statistics Data Analysis, 52(1), 374-393.

[22] Pourahmadi, M. (2013): *High dimensional covariance estimation: with high dimensional data.* Hoboken, New Jersey: John Wiley Sons. E-book.

[23] Team, R. C. (2014): *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

[24] Shen, W., Wang, J. & Ma, S. (2014): *Doubly regularized portfolio with risk minimization.* In Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, 1, 1286-1292. AAAI Press.

[25] Tibshirani, R. (1996): *Regression Shrinkage and Selection via the Lasso.* Journal of the Royal Statistical Society. Series B (Methodological), 58(1), 267-288.

[26] Zhou, D. X. (2013): *On grouping effect of elastic net.* Statistics And Probability Letters, 83(9), 2108-2112.

[27] Zheng, B. & Agresti, A. (2000): *Summarizing the predictive power of a generalized linear model*. Statistics in Medicine, 19(13), 1771-1781.

[28] Zou, H. & Hastie, T. 2005. *Regularization and variable selection via the elastic net.* Journal of the Royal Statistical Society. Series B (Statistical Methodology), 67(2), 301-320.