

VU AMSTERDAM

RESEARCH PAPER IN BUSINESS ANALYTICS

Feature Selection using LASSO

Author:

Valeria FONTI

Supervisor:

Dr. Eduard BELITSER

March 30, 2017



Abstract

Which are the most relevant attributes to describe a response variable?

This is one of the first question a researcher need to ask himself while analyzing a dataset, and the answer is not trivial.

This research paper aims to explain and discuss the use of the LASSO method to address the feature selection task.

Feature selection is a crucial and challenging task in the statistical modeling field, there are many studies that try to optimize and standardize this process for any kind of data, but this is not an easy thing to do.

In the first chapter an introduction of feature selection task and the LASSO method are presented.

In the second chapter we will apply the LASSO feature selection property to a *Linear Regression* problem, and the results of the analysis on a real dataset will be shown.

Finally, in the third chapter the same analysis is repeated on a *Generalized Linear Model* in particular a *Logistic Regression Model* for a high-dimensional dataset. In the same chapter the findings of the scientific study of J.Chen and Z.Chen [4] are presented.

Contents

1	Introduction	3
1.1	Feature Selection	3
1.2	What is LASSO?	4
2	LASSO in Linear Models	6
2.1	Linear Model	6
2.1.1	Statistical Model	6
2.1.2	Linear Model	7
2.2	The LASSO estimator	8
2.2.1	Elastic Net	9
2.2.2	LASSO R implementation	10
2.3	Example: mtcars	11
2.3.1	Dataset description	11
2.3.2	Feature Selection: LASSO	13
3	LASSO in GLM	17
3.1	GLM	17
3.1.1	Logistic Regression Model	18
3.1.2	LASSO Logistic Regression	18
3.2	High-dimensional datasets	19
3.3	Related Work: Chen and Chen's scientific paper	20
3.3.1	EBIC and LASSO	20
3.4	Example: Prostate cancer data	21
3.4.1	Results	21
4	Conclusion	24
	References	25

1 Introduction

In statistics one of the main goals is to build a model that better represent a dataset, this process include the task of features selection.

The only aim of the researcher is to build a model that describes a response variable; in order to do so one of the first question that the researcher should be able to answer is *which features/variables should I take into consideration?* or *Which are the most important attributes to describe the response variable?*.

This research paper aims to answer this questions showing the process of feature selection and describing one of the possible methods to accomplish this task. In particular the focus is on feature selection using the LASSO method.

The goal of this paper is to describe the Lasso method. The LASSO method will be analyzed for both the linear models and the Generalized Linear Models. Furthermore in order to test the efficiency of LASSO, the method will be applied to real data, and the results will be analyzed and described.

1.1 Feature Selection

The feature selection is the process that choose a reduced number of explanatory variable to describe a response variable. The main reasons why feature selection is used are:

- make the model easier to interpret, removing variables that are redundant and do not add any information;
- reduce the size of the problem to enable algorithms to work faster, making it possible to handle with high-dimensional data;
- reduce overfitting.

The variable selection is even more important for the high-dimensional datasets; here the number of features is very high, sometimes higher than the number of observation. In these situation it is hard to easily say which of the variables are relevant and which ones are irrelevant, and on the other hand,

it is difficult, due to dimensionality issues, to build and interpret a model that takes into consideration all the variables. For these reasons the feature selection is an important task.

In the literature there are several types of methods to complete the feature selection task.

First the *Filter Methods* select the features by ranking them on how useful they are for the model, to compute the usefulness score statistical test and correlation results are used (e.g. Chi-square, ANOVA, Pearson's correlation). Secondly *Wrapper Methods* generates different subsets of features, each subset is then used to build a model and train the learning algorithm. The best subset is selected by testing the algorithm. To select the features for the subsets different criteria are used (e.g. Forward and Backward selection)

Finally the *Embedded Methods* are a combination between the two previous methods. The *Embedded Methods* included the LASSO methods that is going to be studied in details in this paper.

1.2 What is LASSO?

LASSO - *Least Absolute Shrinkage and Selection Operator* - was first formulated by Robert Tibshirani in 1996. It is a powerful method that perform two main tasks: regularization and feature selection. The LASSO method puts a constraint on the sum of the absolute values of the model parameters, the sum has to be less than a fixed value (upper bound). In order to do so the method apply a shrinking (regularization) process where it penalizes the coefficients of the regression variables shrinking some of them to zero. During features selection process the variables that still have a non-zero coefficient after the shrinking process are selected to be part of the model. The goal of this process is to minimize the prediction error.

In practice the tuning parameter λ , that controls the strength of the penalty, assume a great importance. Indeed when λ is sufficiently large then coefficients are forced to be exactly equal to zero, this way dimensionality can be reduced. The larger is the parameter λ the more number of coefficients are shrunked to zero. On the other hand if $\lambda = 0$ we have an OLS (Ordinary Least Square) regression.

There are many advantages in using LASSO method, first of all it can provide a very good prediction accuracy, because shrinking and removing the coefficients can reduce variance without a substantial increase of the bias, this is especially useful when you have a small number of observation and a large number of features. In terms of the tuning parameter λ we know that bias increases and variance decreases when λ increases, indeed a trade-off between bias and variance has to be found.

Moreover the LASSO helps to increase the model interpretability by eliminating irrelevant variables that are not associated with the response variable, this way also overfitting is reduced. This is the point where we are more interested in because in this paper the focus is on the feature selection task.

2 LASSO in Linear Models

In this chapter we introduce the linear models together with a brief explanation of their main features. Moreover we explain how to use the LASSO method in this context.

2.1 Linear Model

2.1.1 Statistical Model

A statistical model is a mathematical representation of a real-world problem. The model should summarize and explain the data as close as possible to the reality but it also should be simple and easy to understand and apply. Usually the researcher has some data collected from the real world and his purpose is to build a model on them without losing too much information. A model can be composed by two different type of variables.

Response variable (dependent variable) is the focus of the experiment, it is the output of the model that the researcher wants to investigate on. The response variable can be a single one, *Univariate Models*, or can be multiple, *Multivariate Models*. In this paper we only consider *Univariate Models*.

Explanatory variables (independent variables) are measured or set by the researcher, these are the input of the model. These variables are called explanatory because they explain how the response variable is affected by their changes.

During the modeling process one of the phases is the features selection, that can be done using the LASSO method. In this task the role of the explanatory variables is central, in fact it is really important, in order to build a good model, to choose the right variables that influence the response variable.

In particular in the high-dimensional datasets, where the number of variables is bigger than the number of observations, the selection of features gains greater importance.

The response variable can have different type of probability distributions. When the response variable is normally distributed we consider the *Linear Model* while when the response variable follows different distributions we talk about *Generalized Linear Model*, explained in chapter 3. In this chapter we will consider the *Linear Model*.

2.1.2 Linear Model

Linear Model, often called *Linear Regression Model*, is the model that describes the relationship between response Y_i and explanatory variables X_{ij} . The case of one explanatory variable is called *Simple Linear Regression* while the case with two or more explanatory variables is called *Multiple Linear Regression*. The first assumption of the *Linear Regression Model* is that the response variable is normally distributed

$$Y_i \sim N(\mu_i, \sigma^2) \quad i = 1, \dots, n$$

with mean μ_i and variance σ^2 .

Another assumption is the linearity of the model, that is a linear relationship between the response variable and the explanatory variables.

The Linear model can be expressed as follow

$$Y_i = \beta_0 + x_{i1}\beta_1 + \dots + x_{ik}\beta_k + \epsilon_i \quad i = 1, \dots, n$$

where the parameters $\beta_0, \beta_1, \dots, \beta_k$ are the regression coefficients and k is the number of explanatory variables. Moreover ϵ_i represent the random error, we assume they have 0 mean, constant variance and they are independent.

$$\epsilon \sim N(0, \sigma^2 I)$$

The vector-notation is also used

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

with response vector $\mathbf{Y}_{n \times 1}$, design matrix $\mathbf{X}_{n \times k}$, coefficient vector $\beta_{k \times 1}$ and the error vector $\epsilon_{n \times 1}$

The goal of linear regression is to fit a straight line to a number of points

minimizing the sum of squared residuals.

Regression models are used for many purposes, for instance analysis of variance (*ANOVA*), parameters estimation, prediction and variable selection. Later, in the practical example, we will show how to implement some of them and we will focus on variable selection using *LASSO method*.

2.2 The LASSO estimator

LASSO is a regularization and variable selection method for statistical models. We first introduce this method for linear regression case. The LASSO minimizes the sum of squared errors, with a upper bound on the sum of the absolute values of the model parameters.

There are different mathematical form to introduce this topic, we will refer to the formulation used by Bühlmann and van de Geer [1].

The lasso estimate is defined by the solution to the l_1 optimization problem

$$\text{minimize } \left(\frac{\|\mathbf{Y} - \mathbf{X}\beta\|_2^2}{n} \right) \quad \text{subject to } \sum_{j=1}^k \|\beta\|_1 < t$$

where t is the upper bound for the sum of the coefficients. This optimization problem is equivalent to the parameter estimation that follows

$$\hat{\beta}(\lambda) = \underset{\beta}{\operatorname{argmin}} \left(\frac{\|\mathbf{Y} - \mathbf{X}\beta\|_2^2}{n} + \lambda \|\beta\|_1 \right)$$

where $\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 = \sum_{i=0}^n (Y_i - (\mathbf{X}\beta)_i)^2$, $\|\beta\|_1 = \sum_{j=1}^k |\beta_j|$ and $\lambda \geq 0$ is the parameter that controls the strength of the penalty, the larger the value of λ , the greater the amount of shrinkage.

The relation between λ and the upper bound t is a reverse relationship. Indeed as t becomes infinity, the problem becomes an ordinary least squares and λ becomes 0. Viceversa as t becomes 0, all coefficients shrink to 0 and λ goes to infinity.

In this research paper we are going to use LASSO for its variable selection property.

When we minimize the optimization problem some coefficients are shrank

to zero, i.e. $\hat{\beta}_j(\lambda) = 0$, for some values of j (depending on the value of the parameter λ). In this way the features with coefficient equal to zero are excluded from the model. For this reason LASSO is a powerful method for feature selection while other methods (e.g. Ridge Regression) are not.

As shown in Figure1 both Ridge Regression and LASSO methods find the first point where the least squares error function border touch the constraint region.

For the LASSO method the constraint region is a diamond, therefore it has corners; this means that if the first point is in proximity of the corner, then it has one coefficient β_j equal to zero. While for the Ridge Regression method the constraint region is a disk, thus it has no corners and the coefficients can not be equal to zero.

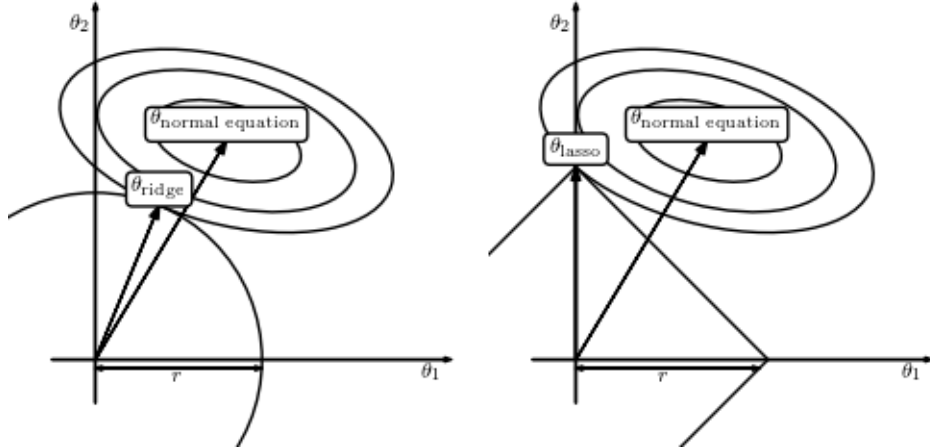


Figure 1: Graph for the Ridge Regression (left side) and LASSO (right side). The two areas are the constraint regions: the disk for the ridge regression, $\beta_1^2 + \beta_2^2 \leq t$, and the diamond for the LASSO, $|\beta_1| + |\beta_2| \leq t$. While the ellipses are the borders of the least squares error functions.

2.2.1 Elastic Net

Elastic Net is an extension of LASSO, it has been proposed by Zou and Hastie in 2005 [7].

The LASSO method has some limitations:

- In small- n -large- p dataset the LASSO selects at most n variables before it saturates.
- If there are grouped variables (highly correlated between each other) LASSO tends to select one variable from each group ignoring the others

Elastic Net overcomes LASSO limitations using a combination of LASSO and Ridge Regression methods.

$$\hat{\beta}_{en}(\lambda) = \underset{\beta}{\operatorname{argmin}} \left(\frac{\|\mathbf{Y} - \mathbf{X}\beta\|_2^2}{n} + \lambda_2 \|\beta\|^2 + \lambda_1 \|\beta\|_1 \right)$$

where $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$ are two regularization parameters.

Adding a quadratic part to the penalty, *Elastic Net* removes the limitation on the selected variables number and stabilize the selection from grouped variables.

2.2.2 LASSO R implementation

In our analysis we will make use of the R statistical software that already has different built in function to implement LASSO. In particular there are two main package that perform regularization methods.

Glmnet (Lasso and elastic-net regularized generalized linear models) is a R package that fits linear models or generalized linear model penalizing the maximum likelihood with both the LASSO method and the Ridge Regression and also the mixture of the two penalties (the elastic net). To find the minimum the glmnet algorithm uses cyclical coordinate descent.

Lars (Least Angle Regression) is a new model-selection method based on the traditional forward selection, i.e. given a collection of possible explanatory variables, we select the one having largest absolute correlation with the response y . In the Lars package there is an implementation of the LASSO method.

The method used to implement the LASSO is similar for both the algorithm. The main difference is that **glmnet** computes a cyclical coordinate descent on a grid of possible value for λ and as output it finds a sequence of models related to the loss function. On the other hand, **Lars** is able to compute the exact value of λ where a new variable enters the model.

However, **glmnet** is the most commonly used nowadays, and one point in its favor is that it can be used for generalized linear model while **Lars** only works for linear regression models. Furthermore there are studies that show that **glmnet** also performs better in many situations.

In conclusion we will use **glmnet** for our analysis because we want to analyze both linear model and generalized linear model.

2.3 Example: mtcars

2.3.1 Dataset description

The dataset contains data extracted from the *Motor Trend* US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 models of car.

There are 32 observation and 10 features, the response variable we want to study is *mpg*, that is the miles per gallon (or fuel efficiency).

The explanatory variables are:

- cyl : Number of cylinders
- disp : Displacement (volume of the engine)
- hp : Gross horsepower
- drat : Rear axle ratio
- wt : Weight (1000 lbs)
- qsec : 1/4 mile time
- vs : V/S engine
- am : Transmission (0 = automatic, 1 = manual)

- gear : Number of forward gears
- carb : Number of carburetors

The goal of our analysis is to underline which explanatory variables are most relevant to predict the response variable *mpg*, in order to do so we will use the LASSO method, but first we will analyze the dataset to understand the data better.

As we can see in Table 1 some correlations between variables are stronger than others. For instance the response variable *mpg* is highly correlated with the following variables: weight (*wt*), number of cylinders (*cyl*), displacement (*disp*) and horsepower (*hp*).

Secondly we can build a linear regression model. One of the example model

Table 1: Correlations : mtcars

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	1										
cyl	-0.85	1									
disp	-0.85	0.9	1								
hp	-0.78	0.83	0.79	1							
drat	0.68	-0.7	-0.71	-0.45	1						
wt	-0.87	0.78	0.89	0.66	-0.71	1					
qsec	0.42	-0.59	-0.43	-0.71	0.09	-0.17	1				
vs	0.66	-0.81	-0.71	-0.72	0.44	-0.55	0.74	1			
am	0.6	-0.52	-0.59	-0.24	0.71	-0.69	-0.23	0.17	1		
gear	0.48	-0.49	-0.56	-0.13	0.7	-0.58	-0.21	0.21	0.79	1	
carb	-0.55	0.53	0.39	0.75	-0.09	0.43	-0.66	-0.57	0.06	0.27	1

to build can be a model that includes the highly correlated explanatory variables. Using the linear regression model we can observe which explanatory variable are more significant for the model and which other are less. For this purpose we will use the **lm** R function and print the ANOVA table.

```
lm<-lm(mpg ~ cyl + disp + hp + wt)
```

Analyzing Table 2 we can see, looking at the p-values, that the variable *cyl*

Table 2: ANOVA					
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cyl	1	817.71	817.71	129.53	0.0000
disp	1	37.59	37.59	5.96	0.0215
hp	1	9.37	9.37	1.48	0.2336
wt	1	90.92	90.92	14.40	0.0008
Residuals	27	170.44	6.31		

and *wt* are very significant; thus we can expect that these variables will be included in the final model.

We still don't know if the choice of the variable included in the model is optimal, probably not because the variable *hp* is not significant and should be excluded from the model.

To decide which variable should be included in the final model we will make use of the LASSO method for variable selection.

2.3.2 Feature Selection: LASSO

The main goal of this research paper is to learn how to use LASSO method for the feature selection task. In the mtcars dataset we want to build a model to predict the response variable *mpg*. To choose which variable to use in the model we use the LASSO method making use of the **glmnet** package in R ($\alpha = 1$ LASSO, $\alpha = 0$ Ridge Regularization).

```
c<-glmnet(as.matrix(mtcars[-1]), mtcars[,1],
           ,standardize=TRUE, alpha=1)
```

Glmnet returns a sequence of different models for different values of λ .

The user has to choose between these models taking into consideration the problem to be solved. We will now use the function on all the explanatory variables in the mtcars dataset, to predict the response variable *mpg*. The results of the analysis are shown in Figure 2 and Figure 3. On the x-axis the different values of λ are shown. Each line represent one of the explanatory variable and its role in the model. In the plots we can see when each variable entered in the model and to which extent they influenced the response variable. Analyzing the plot in Figure 2 we can say that the

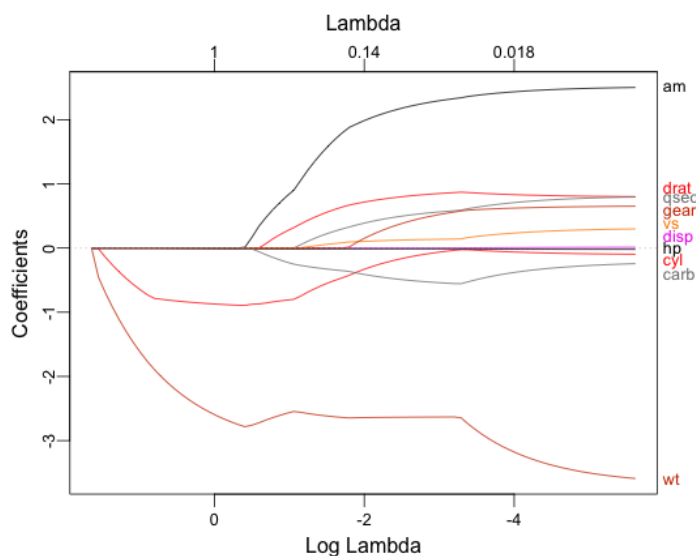


Figure 2: Glmnet: all variables

variable that most influence the model is *wt*, because it enters the model first, and steadily negatively affect the response variable.

The second most important variable is *am* that enters later in the model but positively affect the response variable; as we can see clearly in the plot *am* also affect the trend of the *wt* variable once it enters in the model.

Furthermore we can select other important variables, looking at their trends, such as: *cyl*, *drat* and *carb*. All the other variables seems to be less significant

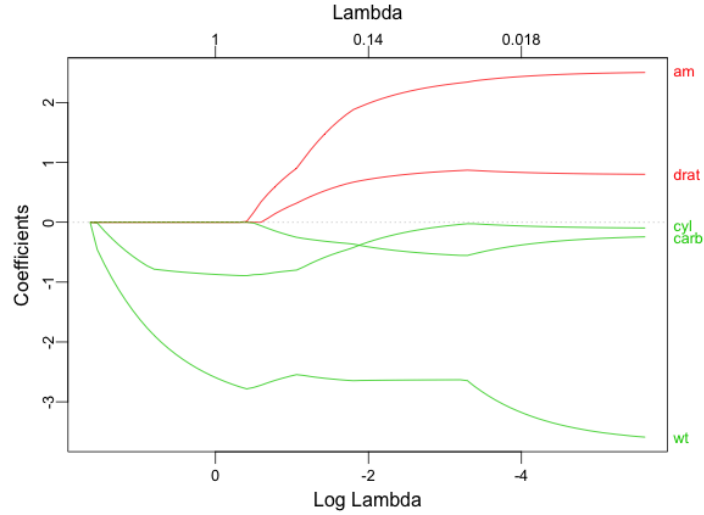


Figure 3: Glmnet: most significant variables

for the model. What has been said can be summarized in Figure 3; where green lines determine the more important attributes that negatively affect the response variable, while red lines determine the ones that affect it positively. The next step is to choose which value of λ to consider. Even if for this small dataset we will not get very informative results, we can still use the *cross-validation*.

```
cv<-cv.glmnet(M[,c(1,3,5)],mtcars[,1],standardize=TRUE,
              ,type.measure='mse',nfolds=5,alpha=1)
```

where M represent a subset of the most important features selected using glmnet.

The R function **cv.glmnet** helps the user to select the most appropriate value for λ , choosing the number of nfold validation. The relevant plot is shown in Figure 4.

In order to choose the most appropriate value for λ the LASSO method

extracts different values for λ , such as λ_{min} (first vertical dotted line) that gives minimum mean cross-validated error and λ_{1se} (second vertical dotted line), that gives a model such that error is within one standard error of the minimum. At this point the user can choose the value for λ that better fit

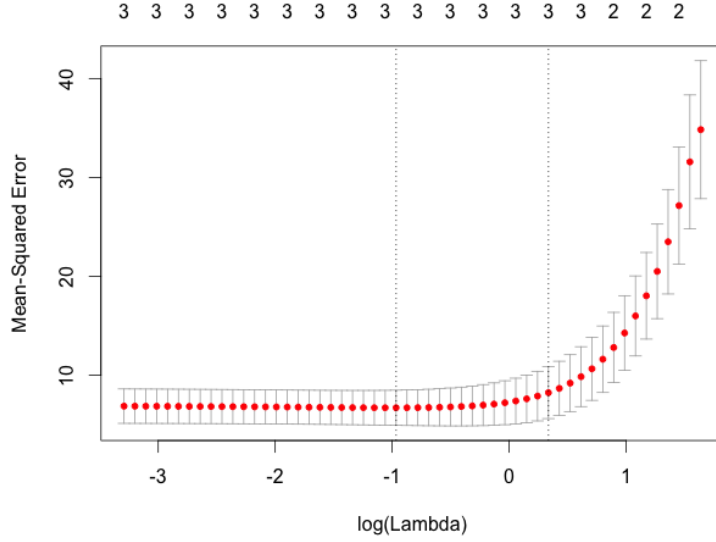


Figure 4: Cross-Validation (nfold=5)

the problem. In our case λ_{min} is not really visible because the plot shows an exponential trend. Moreover for both the values of λ three features are selected and as expected, the most significant variables, shown in Figure 3, are chosen.

For comparison we conduct the same analysis using **Lars** R package and we obtained similar results, the selected variables by **Lars** are *cyl* and *wt*.

```
lars <- lars(as.matrix(mtcars[-1]), mtcars$mpg,
            ,type="lasso")
cv <- lars::cv.lars(as.matrix(mtcars[-1]), mtcars$mpg,
                  ,plot.it = TRUE)
```

3 LASSO in GLM

In this third chapter we briefly introduce the Generalized Linear Models. Furthermore we describe issues and possible solutions with high-dimensional datasets, and we explain which methods are more efficient and how they can be used with high-dimensional dataset. Finally we show how to use LASSO method in the GLM context.

3.1 GLM

Generalized Linear Models are a generalization of the usual *Linear Model* that permits the response variable to have a different distribution than the normal one. The assumption is that the probability distribution has to be from the exponential family, this means that it can be rewrite in the exponential family formulation. There is a wide range of possible distributions for the response variable \mathbf{Y} , such as *Normal*, *Multinomial*, *Poisson* or *Binomial*. Furthermore we have k explanatory variables, \mathbf{X} .

There are three main components in the GLM:

Random Component . The random component states the probability distribution f of the response variable.

$$Y_i \sim f \quad \text{with } Y_i \in \mathbb{R}^n \quad \forall i = 1, \dots, n.$$

Systematic Component . It specifies the linear combination of the explanatory variables, it consist in a vector of predictors η_i

$$\eta_i = x_i^T \beta \quad \text{with } \beta, x_i \in \mathbb{R}^k \quad \forall i = 1, \dots, k.$$

Link Function . It connects the random and the systematic component. It shows how the expected value of the response variable is connected to the linear predictor of explanatory variables.

$$\eta_i = g(\mu_i) \quad \forall i = 1, \dots, n.$$

where μ is the expected value for Y_i and g is a strictly monotone link function.

Table 3: Canonical link functions

Model	Distribution	Link Function
Linear Regression	Normal	Identity
Logistic Regression	Binomial	$\log[\mu_i/(1 - \mu_i)]$
Log-linear	Poisson	$\log(\mu_i)$

There are many different *link functions* existent, the most used are the canonical ones shown in Table 3.

3.1.1 Logistic Regression Model

The *Logistic Regression* is a regression analysis where the response variable is binary, that means it can only assume 0 or 1 values. The explanatory variables can be either discrete or continuous. For this model the main components are:

Random Component . The probability distribution f of the response variable is *Binomial*.

$$Y_i \sim \text{Binomial}(n_i, \pi_i)$$

where n_i is the binomial denominator and π_i is the probability.

Systematic Component , It is the linear combination of the explanatory variables.

$$\eta_i = x_i^T \beta$$

Link Function . The link function is the *logit* function.

$$\eta_i = \text{logit}(\pi) = \log\left(\frac{\pi}{1 - \pi}\right) = \sum_{j=0}^p \beta_j x^{(j)}$$

3.1.2 LASSO Logistic Regression

The LASSO method puts a fixed upper bound on the sum of the absolute values of the model parameters. In the GLM case this constraint can be

expressed by penalizing the negative log-likelihood with l_1 -norm.

In the *Logistic regression* model the negative log-likelihood is given by

$$-\sum_{i=1}^n \log(P_{\beta}(Y_i|X_i)) = \sum_{i=1}^n \left\{ -Y_i \left(\sum_{j=0}^p \beta_j x^{(j)} \right) + \log(1 + \exp(\sum_{j=0}^p \beta_j x^{(j)})) \right\}.$$

That can be written in terms of the loss function ρ :

$$\rho_{(\beta)}(x, y) = -y \left(\sum_{j=0}^p \beta_j x^{(j)} \right) + \log(1 + \exp(\sum_{j=0}^p \beta_j x^{(j)})).$$

The LASSO estimator for a *Logistic regression* model is defined as:

$$\hat{\beta}(\lambda) = \underset{\beta}{\operatorname{argmin}} (n^{-1} \sum_{i=1}^n \rho_{(\beta)}(X_i, Y_i) + \lambda \|\beta\|_1).$$

3.2 High-dimensional datasets

Before introducing the practical example for LASSO in GLM, we want first to mention the typical problems and dataset for which the feature selection task is a crucial part of the analysis.

The high-dimensional datasets, also called small-n-large-p datasets ($p \gg n$), are one of the biggest challenges for researchers. In this type of dataset the number of features is much larger than the sample size, and usually data are sparse, this means that only a few features really affect the response variable. For these two reasons feature selection here is task on which researchers focus most.

Another task, closely related to the feature selection, is the model selection. This task consist in a selection of candidate models and then a criterion to choose one model among the candidates proposed. There are many traditional criteria to address this task, for instance the AIC (Akaike's information criterion) and the BIC (Bayesian information criterion), but many of them fail when dealing with small-n-large-P datasets.

In the next sections we are going to analyze and review the paper from Chen and Chen [4] that propose a solution for the model selection task in high-dimensional datasets. Moreover we will implement the LASSO method on a *Logistic regression* model in a high-dimensional environment.

3.3 Related Work: Chen and Chen’s scientific paper

The second goal of this paper is to present and discuss the contents of the scientific article by J. Chen and Z. Chen ”*Extended BIC for small-n-large-P sparse GLM*” [4]. The paper shows different approaches to deal with small-n-large-P datasets, in particular a new model selection criterion is presented.

3.3.1 EBIC and LASSO

The EBIC, *Extended Bayesian Information Criterion*, is an extension that goes beyond the traditional BIC model selection criterion.

BIC is a criterion that introduce a penalty on the number of parameters in the model and approximates the posterior probability when the prior is uniform on the model. The main limitation of this method is encountered in high-dimensional datasets, where BIC fails and often returns too many variables in the selected model.

The extended BIC (EBIC) takes into account the complexity of the data and in the paper it is demonstrated that, for high-dimensional dataset, it gives very good results.

We will not go into the details of the EBIC formulation because it is not the goal of our paper.

Basically each model selected among the model candidates is created by using a feature selection method, in this case we will describe how to address the feature selection task using the LASSO method.

The show in practice how to implement LASSO on high-dimensional GLM we will use the same example as in the paper we are reviewing [4]. The model used is the *Logistic Regression Model* with logit link function, and it will be implemented with the R package **glmnet**.

3.4 Example: Prostate cancer data

The data we used are taken from the *Singh et al. (2002)* study.

Singh et al. (2002) is a built-in R dataset, that can be loaded through the "sda" package.

The dataset contains prostate cancer tumor data, in particular 6033 genes expression data. The sample size counts 102 observations (men), whereof 52 have the prostate tumor and 50 does not have the prostate tumor.

The response variable Y is a binary variable that shows the presence or absence of the prostate tumor (*cancer* or *healthy*).

The goal of the study is to build a model that correctly classifies the cancer and the healthy observations.

As we can see here we deal with a $p \gg n$ situation, thus probably not all the 6033 genes expressions are relevant for the problem the study wants to solve. The purpose of our analysis is to address the feature selection task and underline which genes are more relevant to predict the response variable, in order to do so we will use the LASSO method.

3.4.1 Results

The dataset has been analyzed using the R **glmnet** function,

```
fit <- glmnet(x, y, family = "binomial",  
             , pmax=20, alpha=1)
```

here x indicates the matrix of the explanatory variables and y is the response variable, $pmax = 20$ means that we want to have maximum 20 genes expressions as result, $alpha = 1$ indicates that we are using the LASSO regularization and $family = "binomial"$ is stated for the *Logistic regression* model.

The result in Figure 5 shows the trends of the 20 most relevant features. Looking at the graph we can clearly see that genes V610 and V1720 influenced the response variable the most and these are also the first two features to enter the model. Thus we can assume that these two genes expressions

are relevant in order to assess if a new sample has the prostate tumor or it does not have.

As for linear regression, the next step would be to find the most appropriate values for λ in order to build a good model.

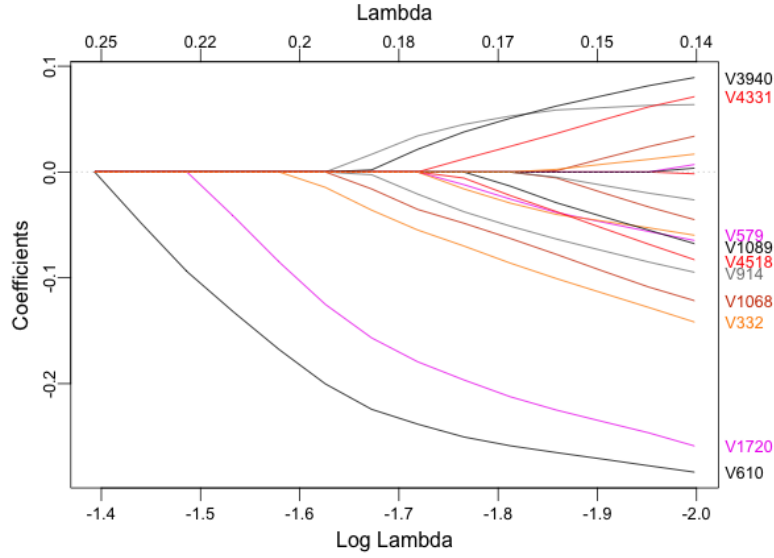


Figure 5: Glmnet: 20 genes

To obtain the values of λ we can analyze the output using the cross-validation, as we did for the linear model in section 2.3.2. In Figure 6 the two vertical dotted lines indicates the two selected values for λ , in particular those are λ_{min} that gives minimum mean cross-validated error or λ_{1se} , that gives a model such that error is within one standard error of the minimum.

```
cvfit<-cv.glmnet(M, y, family = "binomial",
                 ,nfolds=5, type.measure='class ', alpha=.99)
```

where M represent the matrix of the features selected by **glmnet**.

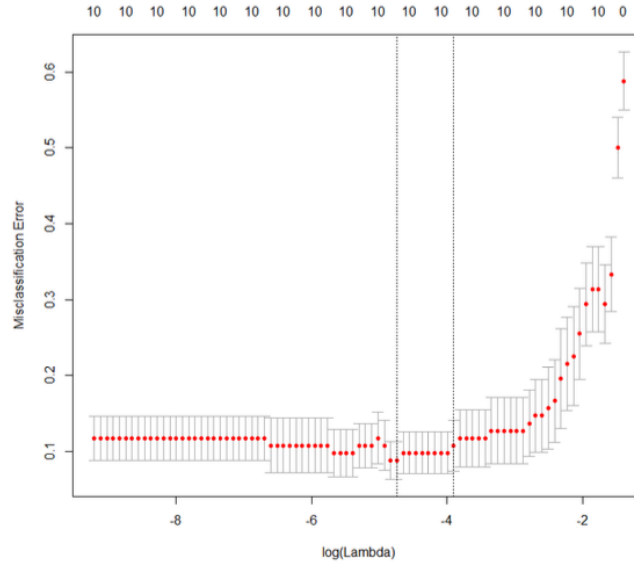


Figure 6: Cross-Validation (nfold=5)

For both the values of λ_{min} and λ_{1se} the following 10 features (genes expressions) are selected: V332, V610, V1720, V364, V1077, V914, V1068, V579, V4331, V3940.

Using this analysis we obtained the most relevant genes in the detection of a prostate cancer.

4 Conclusion

The goal of this research paper was to explain and show how to deal with one of the main statistical modeling task: feature selection.

Feature selection is crucial and challenging in this field of study, mainly because the desired output varies for different set of data, and it is hard to find a model that works for every kind of problem. For these reasons researchers always try to find feature selection model that are well adaptable for the dataset they want to analyze. The task becomes even more challenging when dealing with high-dimensional datasets.

In this paper we decided to face the feature selection problem using the LASSO method. We tested this method using different setups, mainly we focused on two type of statistical models: *Linear model*, *Generalized linear model*. For the GLM we considered the *Logistic regression model* for a small-n-large-P dataset.

Finally, we can say that in both our examples the LASSO method helps us to choose a model with the most relevant features in it.

Further improvement are possible, first Elastic Net can be used to overcome LASSO's limitations.

Secondly, a selection of the most appropriate model can be done using information criteria, such as EBIC [4]. These criteria can be used to evaluate different models using different values of λ .

References

- [1] P.Bühlmann, S.van de Geer: *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer 2011
- [2] Matthew Shardlow: *An Analysis of Feature Selection Techniques*.
- [3] M.C.M. de Gunst: *Statistical Models*. 2013
- [4] J.Chen, Z.Chen: *Extended BIC for small-n-large-P sparse GLM*.
University of British Columbia and National University of Singapore 2012
- [5] Alvin C. Rencher, G. Bruce Schaalje: *Linear Models in Statistics*.
Department of Statistics, Brigham Young University, Provo, Utah 2008
- [6] Robert Tibshirani: *Regression Shrinkage and Selection via the Lasso*.
Journal of the Royal Statistical Society B, Vol.58, Issue 1, pag. 267-288,
1996.
- [7] Hui Zou and Trevor Hastie: *Regularization and variable selection via the elastic net*.
Stanford University, USA, 2005.
- [8] T. Hastie, R. Tibshirani and J. Friedman: *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*.
Springer, 2008
- [9] https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html#lin