Features selection

Small-n large-P sparse GLM: case study

Adnan Kuhait

Research Paper

Master of Business Analytics

Faculty of Sciences

VU University Amsterdam

Supervisor: Dr. E.N. Belitser

February

2016

Abstract

In this paper we will review the paper titled “Extended BIC for small-n-large-P sparse glm” (Chen, Jiahua and Zehua, 2012). The situation of small-n-large-P has become common in genetics research, medical studies, risk management, and other fields. Feature selection is crucial in these studies yet poses a serious challenge. The traditional criteria such as AIC, BIC, and cross-validation choose too many features. In this paper, EBIC is shown to be variable selection consistent under generalized linear models.

First we will introduce the linear model then we will show the needs for the generalized linear model with some examples. Feature and model selection is the title of what comes next, then how to prepare data and what are the conditions to use the method explained in this paper. $glmnet$ in R is the package used to solve the problem, then we will explain the method using two examples of real data.

Preface

Writing a research paper is part of acquiring the Master's degree in Business Analytics at the VU University Amsterdam. The purpose of this paper is to give the student the opportunity to gain experience in doing research on a topic of interest and use the techniques and knowledge that the student has obtained during the study.

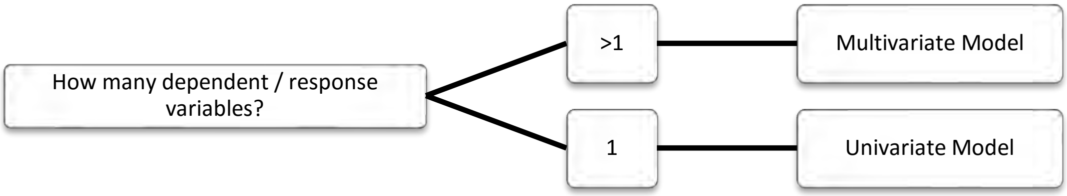
The subject of this paper is to review the paper titled "EXTENDED BIC FOR SMALL-N-LARGE-P SPARSE GLM" (Chen, Jiahua and Zehua, 2012).

My supervisor during this research was *Dr. E.N. Belitser* of the VU University. I want to thank him for all the help and support.

Introduction

The most of the statistical methods involve the analysis of relationships between measurements made on groups of subjects or objects. Measurements consist of two types of variables: explanatory and response variables. In order to analyze data we need first to model these data, the structural form of the model gives better look to the patterns of interactions or associations in data. Inference for the model parameters can show how and which explanatory variable(s) are related to the response variable(s). These variables can be measured in different scales such as Nominal classifications, Ordinal classifications or Continuous measurements.

A model with multiple response variables modeled jointly is called *Multivariate Model* and if the model contains one response variable then is called *Univariate Model*.



These models (multiple or univariate) could be linear or non-linear models. A model is linear when each term is either a constant or the product of a parameter and a predictive variable; a linear equation can then be constructed by adding the results for each term. If the equation doesn't meet the criteria above for a linear equation, it's nonlinear.

The response variable in these models could be normally distributed which is known as *Linear Model* whilst the *Generalized Linear Model* is an extension of the linear model that allows the specification of models whose response variable follows different distributions.

On the other hand, a common misunderstanding is between the words multiple and multivariate. While the word "multiple" applies to the number of predictors that enter the model (or equivalently the design matrix) with a single outcome (Y response), the word "multivariate" refers to a matrix of response vectors (how many dependent variables / outcomes you have).

Here we will give a short overview of Generalized Linear Models (GLM). We shall see that these models extend the linear modeling framework to dependent (or responses) variables that are not normally distributed.

Linear Models

Before we delve into the GLM model, we will take a fast look at the Linear Model. Linear models are those statistical models in which a series of parameters are arranged as a linear combination of the parameters which describe the model under consideration. The term 'linear' in this context does not pertain to the nature of the relationship between the response variable and the predictor variable(s) but rather to the linear relation with respect to the parameter.

Assuming that we have a response variable (Y) and explanatory variables (X's), then we can model the responses as a function of the explanatory variables:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \varepsilon_i.$$

The response Y_i , i = 1,. . . ,n is modeled by a linear function of explanatory variables x_j , j = 1, . . . , p plus an error term.

This can be written in a more general form

$$Y = X\beta + \varepsilon.$$

Where

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

Y is the $n\ x\ 1$ vector of expectations, and X is so called the design matrix[1]. Errors ε_i are assumed independent and identically distributed such that:

$$E(\varepsilon_i) = 0, and\ \ var(\varepsilon_i) = \sigma^2.$$

Some examples of the application of the linear models are: Simple linear regression, multiple regression, one-way ANOVA and two-way ANOVA.

A regression with two or more explanatory variables is called a multiple regression (the term was first used by Pearson, 1908). Rather than modeling the mean response as a straight line, as in simple regression, it is now modeled as a function of several explanatory variables. The general purpose of multiple regression is to analyze the relationship between several independent or predictor variables and a dependent or criterion variable.

The computational problem that needs to be solved in multiple regression analysis is to fit a straight line (or plane in an n-dimensional space, where n is the number of independent variables) to a number of points. In the simplest case - one dependent and one independent variable - we can visualize this in a scatterplot (scatterplots are two-dimensional plots of the scores on a pair of variables).

[1] Is a matrix of values of explanatory variables, often denoted by$X$.

The function ***lm*** can be used to perform multiple linear regressions in R and much of the syntax is the same as that used for fitting simple linear regression models.

$$lm(response \sim explanatory\_1\ +\ explanatory\_2\ +\ \dots +\ explanatory\_p)$$

R uses

+ To combine elementary terms, as in $A + B$.

: For interactions, as in$A{:}\,B$;

* For both main effects and interactions, so $A * B\ =\ A + B + A{:}\,B$.

Here the terms $response$ and $explanatory\_i$ in the function should be replaced by the names of the response and explanatory variables, respectively, used in the analysis.

In general, the purpose of analysis of variance (ANOVA) is to test for significant differences between means. Elementary Concepts provides a brief introduction to the basics of statistical significance testing. If we are only comparing two means, $ANOVA$ will produce the same results as the $t - test$ for independent samples (if we are comparing two different groups of cases or observations) or the t test for dependent samples (if we are comparing two variables in one set of cases or observations).

Example[2]:

A data file containing information on three variables for 20 countries in Latin America:

	Setting	Effort	Change
Bolivia	46	0	1
Brazil	74	0	10
Chile	89	16	29
Colombia	77	16	25
Costa Rica	84	21	29
Cuba	89	15	40
Dominican Rep	68	14	21
Ecuador	70	6	0
El Salvador	60	13	13
Guatemala	55	9	4
Haiti	35	3	0
Honduras	51	7	7
Jamaica	87	23	21
Mexico	83	4	9
Nicaragua	68	0	7
Panama	84	19	22
Paraguay	74	3	6
Peru	73	0	2
Trinidad Tobago	84	15	29
Venezuela	91	7	11

Table 1 information on three variables for 20 countries in Latin America

[2] http://data.princeton.edu/wws509/datasets/effort.dat

This small dataset includes an index of social setting, an index of family planning effort, and the percent decline in the crude birth rate between 1965 and 1975.

To fit an ordinary linear model with setting and effort as predictors and change as the response variable, we will use the following model:

```
lmfit  < − lm(change~setting + effort)

> lmfit

 Call:

lm(formula  =  change ~ setting  +  effort)

 Coefficients:

(Intercept)    setting     effort

  −14.4511      0.2706      0.9677
```

The output includes the model formula and the coefficients. To get a hierarchical analysis of variance table corresponding to introducing each of the terms in the model one at a time, in the same order as in the model formula, try the *anova* function:

```
> anova(lmfit)

Analysis of Variance Table

Response: change

Df  Sum Sq Mean Sq F value   Pr(> F)

setting    1 1201.08 1201.08  29.421 4.557e − 05  ***

effort     1  755.12  755.12  18.497 0.0004841  ***

Residuals 17  694.01   40.82

− − −

Signif. codes:  0  `***′  0.001  `**′  0.01  `*′  0.05  `.′  0.1  `′  1
```

Alternatively, we can plot the results using

```
> plot(lmfit)
```

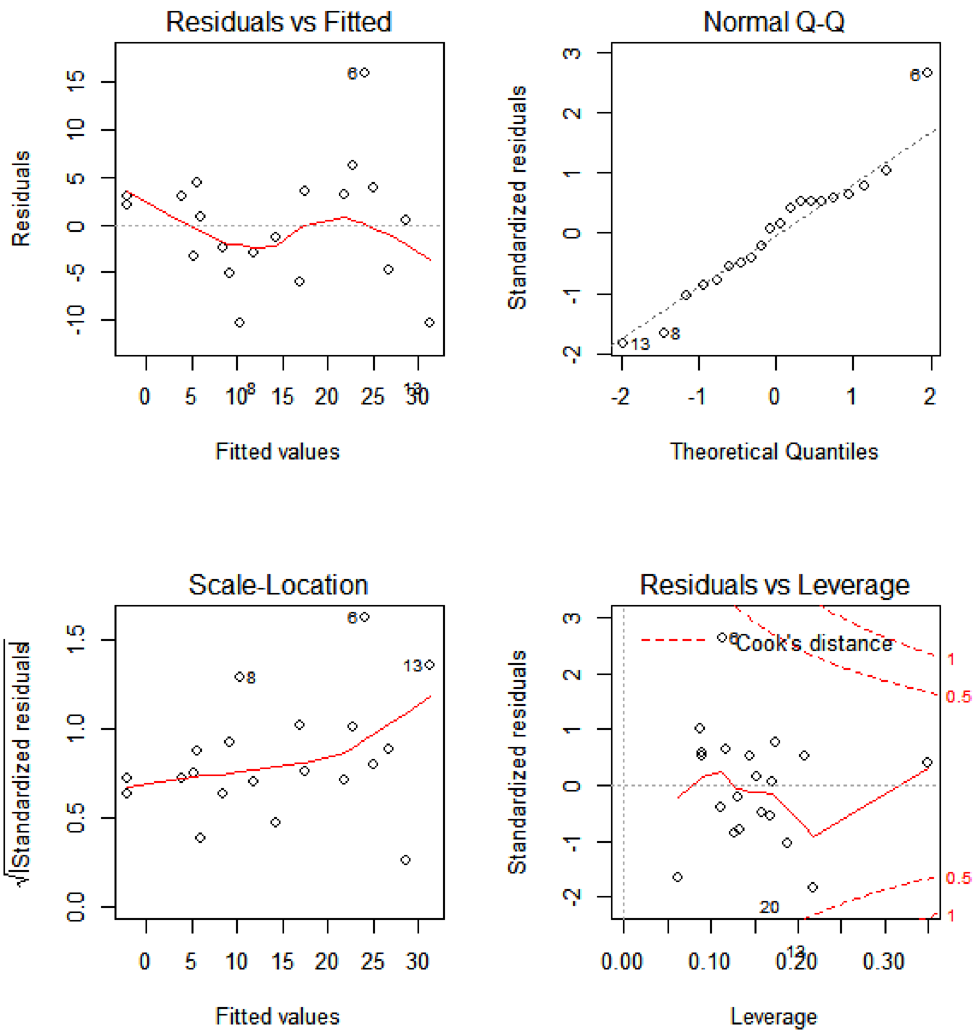


Figure 1 plot(lmfit)

Furthermore, we can extract much information from our $lmfit$ model such as:

$fitted\ (lmfit)$	: extracts the fitted values.
$coef\ \ (lmfit)$	: To extract the coefficients.
$residuals(lmfit)$	: to get the residuals.

Generalized Linear Model

The ordinary linear models are, in fact, special cases of generalized linear models (GLMs). Both generalized linear models and ordinary linear models investigate the relationship between a response variable and one or more predictors. Both techniques estimate parameters in the model so that the fit of the model is optimized. But what if the data follows probability distributions other than the Normal distribution, such as the Poisson, Binomial, Multinomial, and etc.? GLM can deal with such situation; also, the GLM can be used for a non-linear relation between the expected responses and the parameters of the model. GLM's include a link function that relates the mean of the response to the linear predictors in the model. In a GLM model there are three components:

1. The random component (the outcome), specifies the distribution of Y_i which it was assumed to be normally distributed in the ordinary linear model,

$$Y_i \sim f_i.$$

2. Systematic component (the design matrix multiplied by the parameter vector), it specifies the way in which the explanatory variables come into the model.

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

3. Link function, denoted by$g(.)$, a function that links the systematic component to the random component.

$$\eta_i = g(\mu_i).$$

Where$\eta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$.

The first two components come from the fact that statistical models contain both systematic effect and random effect, while the third component is what links them together.

Assuming that the observations come from a distribution in the exponential family with probability density function (or probability mass function):

$$f_i(y) = f_i(y, \theta_i) = \exp\left(\frac{y\theta_i - b(\theta_i)}{\phi/A_i} + c\big(y, \ \phi/A_i\big)\right).$$

Where

- b is an arbitrary monotonic, differentiable function.
- θ_i is the one parameter of the exponential family specific to Y_i .
- ϕ is a (possibly known) scale parameter.
- $A_i$denotes a known weight constant.

If $Y_i$has a distribution in the exponential family then it has mean and variance:

$$E(Y_i) = \mu_i = b'(\theta_i)\,, \qquad var(Y_i) = b''(\theta_i)\,\phi/A_i\,, \qquad \text{For i = 1,...,n.}$$

Where $b'(\theta_i)$ and $b''(\theta_i)$ are the first and second derivatives of$b(\theta_i)$. The link function g describes how the mean, $E(Y_i) = \mu_i$ depends on the linear predictor here g expresses $\eta_i$as a function of$E(Y_i)$.

Examples of Generalized Linear Models

You construct a generalized linear model by deciding on response and explanatory variables for your data and choosing an appropriate link function and response probability distribution. Explanatory variables can be any combination of continuous variables, classification variables, and interactions.

Logistic regression

In the classical framework, we were interested in modeling a continuous response variable y as a function of one or more predictor variables.

Example: *Modeling Binomial Data*

Suppose $Y_i \sim Binomial\ (n_i, p_i)$.

And we wish to model the proportions Y_i/n_i , then

$$E(Y_i/n_i) = p_i,$$

$$var(Y_i/n_i) = \frac{1}{n_i} p_i\,(1 - p_i).$$

Link function must map from$(0,1) \rightarrow (-\infty\ , \infty)$. A common choice is:

$$g(\mu_i) = logit\ (\mu_i) = \log(\frac{\mu_i}{1-\mu_i}).$$

Poisson regression

The Poisson probability distribution is perhaps the most commonly used discrete distribution for modeling count data.

Example: *Modeling Poisson Data*

Suppose $Y_i \sim Poisson\ (\lambda_i)$.

Then $E(Y_i) = \lambda_i,$

$$var(Y_i) = \lambda_i.$$

Our link function must map from$(0, \infty) \rightarrow (-\infty\ , \infty)$. A natural choice is

$$g(\mu_i) = log(\mu_i).$$

Feature and model selection

Researchers try to find the relation between explanatory features and a response variable. This includes the challenge of feature selection especially the case of small sample ($small - n$) and extremely large features ($large - P$) and then its sparsity, in which only few unidentified features affect the response variable.

There are a wide range of traditional model selection criteria. However, in the case of small-n-large-P, these criteria often fail to serve the purpose of feature selection. To solve this limitation, scientists propose AIC, BIC, the extended Bayes Information Criteria (EBIC) and other methods. Under the ordinary linear model, EBIC is shown to be selection consistent in the small-n-large-P situation. However, its validity under other regression models is still unsolved.

In the paper we are reviewing "EXTENDED BIC FOR SMALL-N-LARGE-P SPARSE GLM" (Chen and Chen 2012); the researchers used tailor-developed technical results for the exponential family distribution. This is for the purpose of proving the uniform consistency of the maximum likelihood estimates of the coefficients in the linear predictor of all GLM models containing causal features and the selection consistency of EBIC under GLM with canonical links. Under some technical conditions, extended BIC (EBIC) shows variable selection consistency under GLM, as the paper of Chen & Chen (2012) demonstrates.

Because we are here concentrating on the applications of this paper, we will not delve in the technical details which is far from the purpose of this review.

Let χ be the set of all features under consideration. Let s be a subset of χ, $\nu(s)$ the number of features in s, and $\beta(s)$ the vector of the components in β that corresponds to the features in s.

Let β_0 be the unknown true value of the parameters. The components of β_0 other than those in s_0 are zero. Let $x_i(s)$ be the vector of the components of x_i that correspond to $\beta(s)$.

Let $\mathrm{A}_0 = \{ s : s_0 \subset s; \nu(s) \leq K\}$, $A_1 = \{s : s_0 \not\subset s; \nu(s) \leq K\}$.

Th1: Under some technical conditions on the model,

$$\max_{s \in A_0} \left\| \hat{\beta}(s) - \beta_0(s) \right\| = O_P(n^{-\frac{1}{3}}), \quad \text{as} \quad n \to \infty.$$

This theorem gives the uniform consistency of the maximum likelihood estimates of the coefficients in the linear predictor of all generalized linear models (GLM) containing causal features.

There are too more theorems in the paper of Chen & Chen (2012).

- The first theorem shows the selection consistency of EBIC under GLM with canonical links.

- The second theorem shows that the EBIC selects almost surely the model that exhausts all *K* retained features.

Data Preparation

Before starting, one needs to check the technical conditions listed in the paper, which briefly are:

- *No two collinear features exists (typical condition in compressed sensing).*
- *Features are assumed to be standardized.*
- *The square of a feature does not have a severely skewed distribution.*

glmnet in R:

$$glmnet(x, y,\ family,\ alpha,\ pmax, \dots)$$

Fit a generalized linear model via penalized maximum likelihood.

$$-loglik/nobs\ +\ \lambda\ *\ penalty$$

The regularization path is computed for the LASSO or elastic net penalty at a grid of values for the regularization parameter lambda.

LASSO: is a shrinkage and selection method for linear regression. It minimizes the usual sum of squared errors, with a bound on the sum of the absolute values of the coefficients.

Choosing alpha in the model is the way to determine which penalty family we will use:

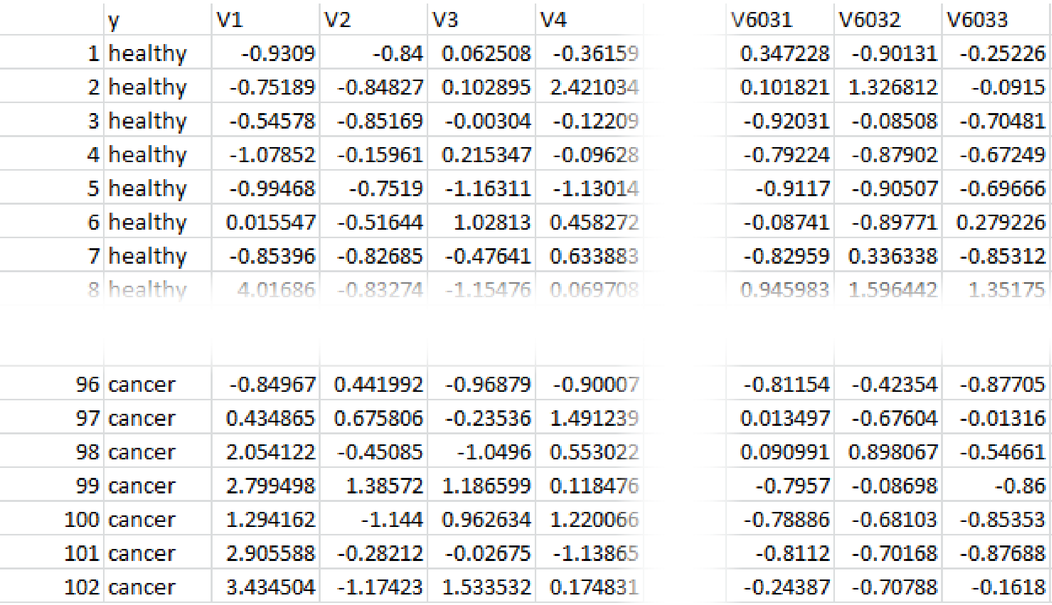
$$(1\ -\ \alpha)/2||\beta||_2^2\ +\ \alpha||\beta||_1$$

- $\alpha = 0$ Ridge Regression which shrinks correlated variables toward each other.
- $\alpha = 1$ LASSO does feature selection.
- $0 < \alpha < 1$ Elastic net can deal with grouped variables.

The authors choose to use Elastic net to obtain regression models with various levels of sparsity.

Example 1: Prostate cancer data set Singh et al. 2002

To show the importance of this method, we used the same data used in the paper which is the *R* pre-loaded dataset *Singh et al. (2002)*. The dataset contains high-quality expression profiles were successfully derived from 52 prostate tumors and 50 healthy prostate samples from patients undergoing surgery. The goal here is to classify tumor and healthy samples. The number of gene expression levels is 6033 genes as shown in Figure 2.



	y	V1	V2	V3	V4	V6031	V6032	V6033
1	healthy	-0.9309	-0.84	0.062508	-0.36159	0.347228	-0.90131	-0.25226
2	healthy	-0.75189	-0.84827	0.102895	2.421034	0.101821	1.326812	-0.0915
3	healthy	-0.54578	-0.85169	-0.00304	-0.12209	-0.92031	-0.08508	-0.70481
4	healthy	-1.07852	-0.15961	0.215347	-0.09628	-0.79224	-0.87902	-0.67249
5	healthy	-0.99468	-0.7519	-1.16311	-1.13014	-0.9117	-0.90507	-0.69666
6	healthy	0.015547	-0.51644	1.02813	0.458272	-0.08741	-0.89771	0.279226
7	healthy	-0.85396	-0.82685	-0.47641	0.633883	-0.82959	0.336338	-0.85312
8	healthy	4.01686	-0.83274	-1.15476	0.069708	0.945983	1.596442	1.35175
96	cancer	-0.84967	0.441992	-0.96879	-0.90007	-0.81154	-0.42354	-0.87705
97	cancer	0.434865	0.675806	-0.23536	1.491239	0.013497	-0.67604	-0.01316
98	cancer	2.054122	-0.45085	-1.0496	0.553022	0.090991	0.898067	-0.54661
99	cancer	2.799498	1.38572	1.186599	0.118476	-0.7957	-0.08698	-0.86
100	cancer	1.294162	-1.144	0.962634	1.220066	-0.78886	-0.68103	-0.85353
101	cancer	2.905588	-0.28212	-0.02675	-1.13865	-0.8112	-0.70168	-0.87688
102	cancer	3.434504	-1.17423	1.533532	0.174831	-0.24387	-0.70788	-0.1618

Figure 2 Prostate cancer data set Singh et al. 2002

The dataset can be loaded after installing the required library in$R$:

```
# load sda library
library("sda")
# load Singh et al (2002) data set
data(singh2002)
```

The first step is to examine the correlation of the gene data, then by choosing $K = 20$ and $\gamma = 0.5$ for the EBIC. Using $glmnet$ in *R*, we were able to identify the most ten effective features among the *6,033* feature.

```
output = glmnet(x, y, family = "binomial", alpha = 0.99, pmax = K)
```

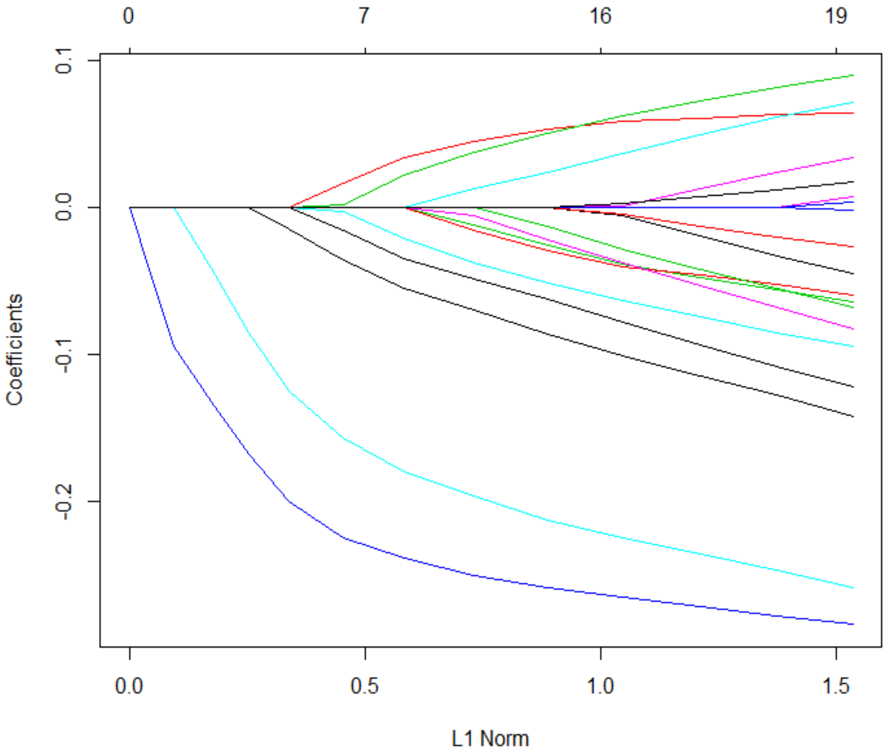


Figure 3 visualization of the coefficients

Figure 3: shows the visualization of the coefficients, each curve corresponds to a variable. It shows the path of its coefficient against the $\ell 1$-norm of the whole coefficient vector at as λ varies. The axis above indicates the number of nonzero coefficients at the currentλ, which is the effective degrees of freedom (df) for the LASSO.

Next step we examine the output using cross validation of $glmnet$ (included in the package)

$$cvfit \ <-cv.glmnet(x[,aa[[k]]\ ],as.numeric(y),family \ = \ "binomial",nfolds \ = \ 5,type.measure \\ = \ "class",alpha \ = \ 0.99)$$

When plotting (Figure 4) the $cvfit$ cross validation model (red dotted line), we can see the upper and lower standard deviation curves along the λ sequence (error bars). Two selected λ's are indicated by the vertical dotted lines, $lambda.min$ is the value of λ that gives minimum mean cross-validated error, and $lambda.1se$ which gives the most regularized model such that error is within one standard error of the minimum.

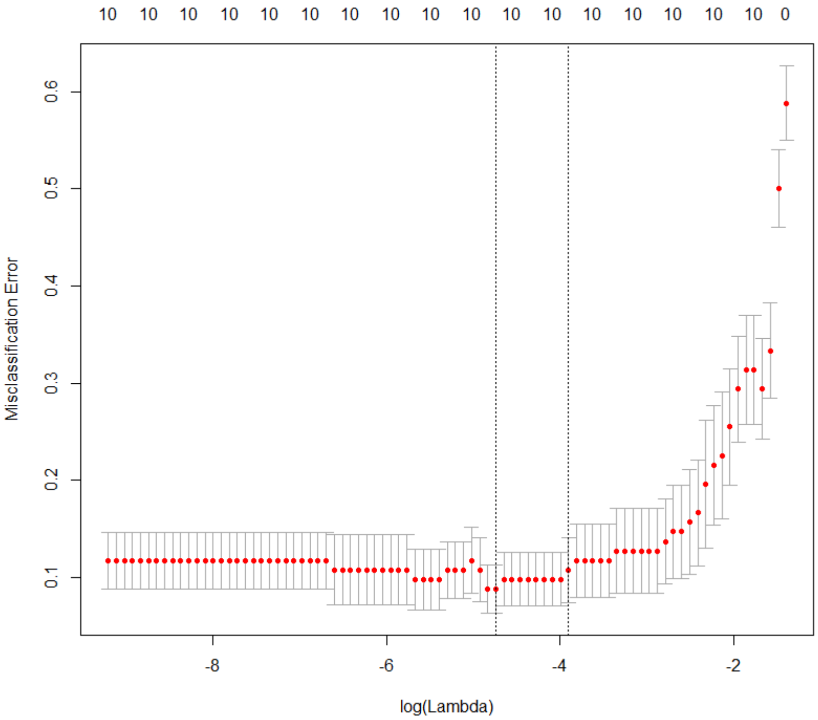


Figure 4 Cross validation

Using $EBIC_{\gamma=0.5}$ we were able to identify the most ten effective features among the *6,033* feature, last step is to order them in importance using $glmpath$.

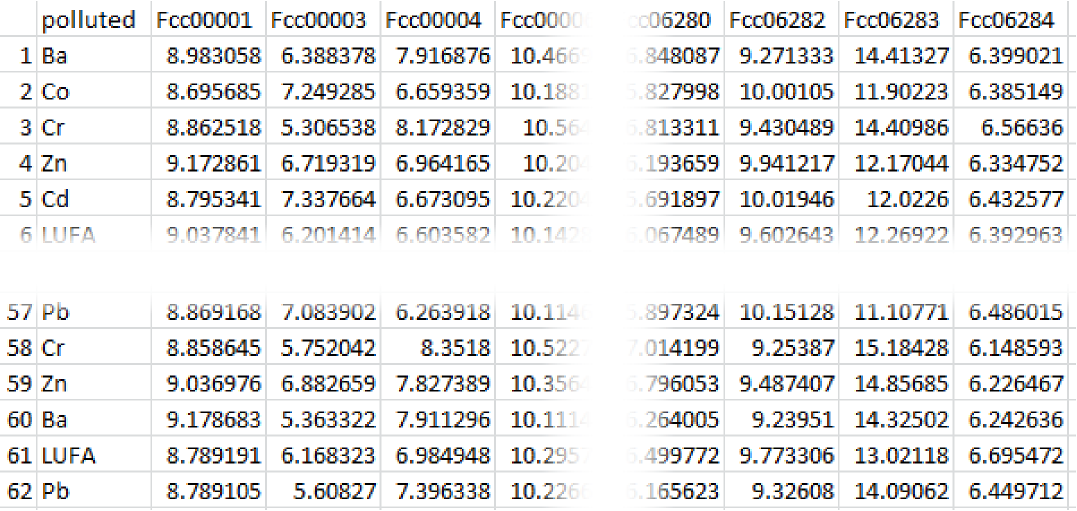
```
gpath <- glmpath(xx[, aa[[k]]], as.numeric(y), family = "binomial", min.lambda = cvfit$lambda.1se)
> gpath
Call:
glmpath(x = xx[, aa[[k]]], y = as.numeric(y), family = "binomial",
  min.lambda = cvfit$lambda.1se)
Step 1 :  V610
Step 2 :  V1720 V332 V364 V1068 V914 V3940 V1077 V4331 V579
```

This will give us the best order according to most effective gene:

Gene No.	610	1720	332	364	1068	914	3940	1077	4331	579

Example 2: Detecting heavy metal pollution in soils

In this example we used a dataset from a research was conducted by the VU University and published in 2010. Scientists would like to use the genes in “Folsomia”, which is a small millimeter–size insect that lives in the soil, to detect heavy metal pollution in soils. Data contains small-n=62 observations and large-P=5070 genes.



	polluted	Fcc00001	Fcc00003	Fcc00004	Fcc0000	c06280	Fcc06282	Fcc06283	Fcc06284
1	Ba	8.983058	6.388378	7.916876	10.466	.848087	9.271333	14.41327	6.399021
2	Co	8.695685	7.249285	6.659359	10.188	.827998	10.00105	11.90223	6.385149
3	Cr	8.862518	5.306538	8.172829	10.56	.813311	9.430489	14.40986	6.56636
4	Zn	9.172861	6.719319	6.964165	10.20	.193659	9.941217	12.17044	6.334752
5	Cd	8.795341	7.337664	6.673095	10.220	.691897	10.01946	12.0226	6.432577
6	LUFA	9.037841	6.201414	6.603582	10.142	.067489	9.602643	12.26922	6.392963
57	Pb	8.869168	7.083902	6.263918	10.114	.897324	10.15128	11.10771	6.486015
58	Cr	8.858645	5.752042	8.3518	10.522	.014199	9.25387	15.18428	6.148593
59	Zn	9.036976	6.882659	7.827389	10.356	.796053	9.487407	14.85685	6.226467
60	Ba	9.178683	5.363322	7.911296	10.111	.264005	9.23951	14.32502	6.242636
61	LUFA	8.789191	6.168323	6.984948	10.295	.499772	9.773306	13.02118	6.695472
62	Pb	8.789105	5.60827	7.396338	10.226	.165623	9.32608	14.09062	6.449712

Figure 5: Detecting heavy metal pollution in soils

as it shown in Figure 5, the table is more informative and gives more details about the pollution material, but for the simplicity we will convert it to binomial table (polluted or not).

Applying the same steps in (Example 1); First step is to check the conditions. Then by choosing $K = 20$ and $\gamma = 0.5$ for the $EBIC_\gamma$. Using $glmnet$ in R, and the $EBIC_\gamma$ criterion we were able to identify the most ten effective features among the *5070* features.

In Figure 6, when plotting the $cvfit$ cross validation model (red dotted line), we can see the upper and lower standard deviation curves along the λ sequence (error bars). Two selected λ’s are indicated by the vertical dotted lines, $lambda.min$ is the value of λ that gives minimum mean cross-validated error, and $lambda.1se$ which gives the most regularized model such that error is within one standard error of the minimum.

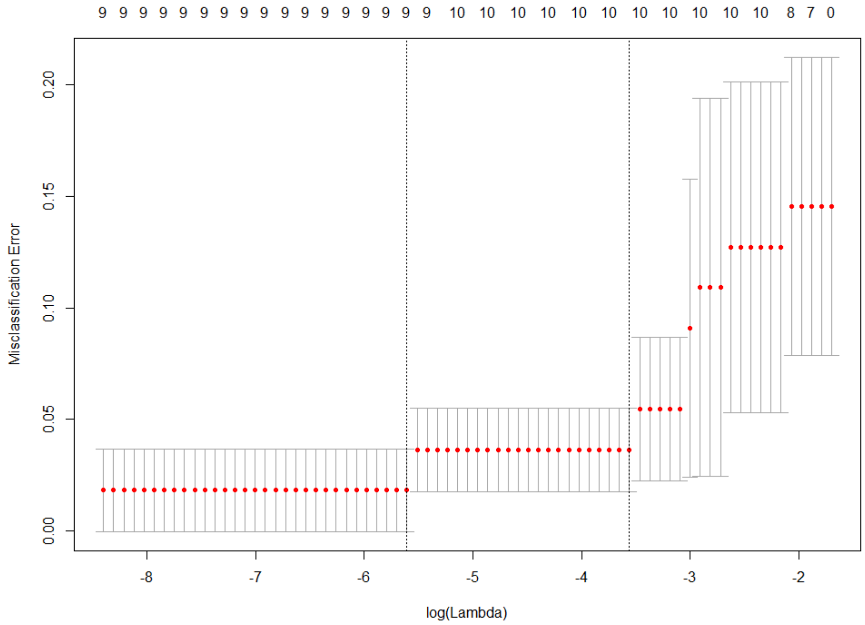


Figure 6 cross validation model

Using *glmpath* to order the result which will give us the following features order:

```
Call:
glmpath(x = x_pol[, aa_pol[[k_pol]]], y = as.numeric(y_pol),
  family = "binomial")
Step 1 :  Fcc01630C1
Step 2 :  Fcc05798
Step 3 :  Fcc03273  Fcc03353C1  Fcc00246C1  Fcc04919  Fcc05730  Fcc00786C1  Fcc01912C1  Fcc02691C1
```

Thus, what shown above are the most effective genes ordered in importance.

Conclusion

The small-n-large-P situation has become common in genetics research, medical studies, risk management, and other fields. Feature selection is crucial in these studies yet poses a serious challenge. The traditional criteria such as AIC, BIC, and cross-validation choose too many features. In the paper of Chen & Chen (2012) which we have reviewed, EBIC is shown to be variable selection consistent under generalized linear models.

In the examples, we examined the variable selection problem under the generalized linear models. Here we used $glmnet$ in R which is a powerful package in this situation. The first example we examined an R pre-loaded dataset, this is the same example used in the paper of Chen & Chen (2012), the reason on why we choose to redo the example is to test and compare our results with the results in the paper, some small details were chanced because of the R versions difference and computers evolution since 2012 till now, but got the same results and we were easily able to identify the most effective genes in the same order importance.

In the second example, we applied the same technique, very simple and notable fast results. The dataset was collected for a research conducted by the Vrije Uneversiteit (VU). Here we were also able to identify the most effective genes with which we can determine the pollution in the soil.

Briefly we can say that the paper succeeded to show an effective way to identify the most powerful features in the case of small-n-large-P.

Bibliography

Chen, Jiahua and Zehua. (2012). EXTENDED BIC FOR SMALL-n-LARGE-P SPARSE GLM. *Statistica Sinica 22*, pp. 555-574.

Dipankar Bandyopadhyay, Introduction to Generalized Linear Models, Analysis of Categorical Data Spring 2011. Medical University of South Carolina

Annette J. Dobson , An Introduction To Generalized Linear Models, Chapman & Hall/CRC, 2002

Heather Turner, Introduction to Generalized Linear Models, Department of Statistics, University of Warwick, UK, 2008.

M.C.M. de Gunst, Statistical Models, VU, 2013

Jong-Hwan Yoo , INTRODUCING THE GENERALIZED LINEAR MODELS

Duncan Anderson et al, A Practitioner's Guide to Generalized Linear Models, Feb 2007

JMP Examples of GLM
(http://www.jmp.com/support/help/Examples_of_Generalized_Linear_Models.shtml)

Benjamin Nota and others, VU University 2010. Gene Expression Analysis Reveals a Gene Set Discriminatory to Different Metals in Soil.

U. Alon et al. (1999): Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissue probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. USA 96, 6745-6750.
Package source: ("https://bioconductor.org/biocLite.R")
biocLite("colonCA")
https://bioconductor.org/packages/release/data/experiment/html/colonCA.html
http://users.monash.edu.au/~murray/stats/BIO4200/LinearModels.pdf

Appendix

BIC: In statistics, the Bayesian information criterion (BIC) or Schwarz criterion (also SBC, SBIC) is a criterion for model selection among a finite set of models; the model with the lowest BIC is preferred. It is based, in part, on the likelihood function and it is closely related to the Akaike information criterion (AIC).

AIC: The Akaike information criterion (AIC) is a measure of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Hence, AIC provides a means for model selection.

Extended BIC: Mathematically is the classical BIC with an additional penalty term *2γ log P* with a positive γ.

***LASSO** :* is a shrinkage and selection method for linear regression. It minimizes the usual sum of squared errors, with a bound on the sum of the absolute values of the coefficients. It has connections to soft-thresholding of wavelet coefficients, forward stagewise regression, and boosting methods.

LASSO is a regularization technique. Use LASSO to:

- Reduce the number of predictors in a regression model.
- Identify important predictors.
- Select among redundant predictors.
- Produce shrinkage estimates with potentially lower predictive errors than ordinary least squares.

Elastic net is a related technique. Use elastic net when you have several highly correlated variables. LASSO provides elastic net regularization when you set the Alpha name-value pair to a number strictly between 0 and 1.

GLMnet in R: Fit a generalized linear model via penalized maximum likelihood. The regularization path is computed for the LASSO or elastic net penalty at a grid of values for the regularization parameter lambda. Can deal with all shapes of data, including very large sparse data matrices. Fits linear, logistic and multinomial, Poisson, and Cox regression models.