A Survey of Classical and Bayesian Model Selection

Math453 Project Report

Lorenzo Croissant

University of Lancaster

l.croissant@lancaster.ac.uk

January 14, 2018

Abstract

The problem of model selection is of paramount importance in any real-world statistical analysis. In particular, in the context of linear regression there are many approaches to performing model selection. In this paper, a selection of common Classical and Bayesian methods are explored and compared. After a presentation of the theoretical framework required, which assumes familiarity with Ordinary Least Squares regression, concepts are used to derive effective Ridge regression selection methods. A simple algorithm is developed and compared with Ridge regression, stepwise selection, and Least Angle regression. Empirical results are discussed using an existing dataset.

Contents

I. Introduction

i. Prior Research

There are many model selection methods for most simple modeling procedures such as linear regression, and many surveys of such methods[1]. Throughout this paper we will look only at empirical applications of selection methods, using the diabetes dataset obtained from the paper introducing Least Angle Regression [2]. It consists of 442 observations of 10 measurements and a response evaluating the progression of diabetes amongst these patients. Before using it, this dataset has been expanded however, to include all square terms and all first order interactions. All covariates have also been normalised by column.

ii. Aims

This paper aims to provide a reader proficient in classical statistics with a simple comparison of different model selection criteria. Theoretical background will thus be provided for Bayesian methods (e.g. evidence) and a simple presentation will be given of the other methods used. Next we will apply the three chosen methods (evidence, stepwise selection, cross-validation) to our dataset. Afterwards we will provide diagnostics and extension methods, before finally offering some closing remarks.

II. Theory

i. Principle of Bayesian Statistics

The key difference between Classical and Bayesian Statistics is the approach to randomness. This is why we will first frame the Bayesian paradigm before examining in detail how the theory is formulated. Once this

is done we will proceed to apply the theory to the case of Multiple Linear Regression (MLR).

When modeling a stochastic situation in Classical (frequentist) statistics, one considers the parameters $\boldsymbol{\theta}$ of the model to be fixed but unknown, and the data $\mathbf{y}$ to be random, taken from a distribution that depends on $\boldsymbol{\theta}$. In general terms, to estimate the value of $\boldsymbol{\theta}$, we would maximise the likelihood function $L(\theta|\mathbf{y}) := f(\mathbf{y}|\boldsymbol{\theta})$, which corresponds to the likelihood that the observed data $\mathbf{y}$ are generated by a given $\boldsymbol{\theta}$ under our model.

As one can see above, frequentists consider the randomness to lie only in the data, not the parameters. Bayesian Statistics, however, rest on the view that parameters should also be considered to have inherent uncertainty instead of being fixed. Bayesian Inference adds three new functions to consider, which we will relate to the likelihood shortly.

First, the prior for $\boldsymbol{\theta}$, $\pi(\boldsymbol{\theta})$ which is the probability distribution of $\boldsymbol{\theta}$ when no data is known. In practice, this is a guess of the behaviour of $\boldsymbol{\theta}$ based on any information besides the data, such as expert advice. Second, the distribution which presents the information about $\boldsymbol{\theta}$ once the data is taken into account is the posterior $\pi(\boldsymbol{\theta}|\mathbf{y})$. Finally the evidence $m(\mathbf{y})$ is an integration constant we will talk more about in subsection iv.

All these quantities are related via Bayes' Theorem:

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{m(\mathbf{y})}$$

This has several noteworthy consequences, such as the posterior being a proper density for $\boldsymbol{\theta}$ (unlike $L(\theta|\mathbf{y})$). Bayes' Theorem also provides a straightforward computation of $\pi(\boldsymbol{\theta}|\mathbf{y})$ up to $m(\mathbf{y})$, which is sufficient to describe its shape and maximum.

ii. Bayesian Regression

Throughout this paper we will consider the following setting: $\mathbf{y} := (y_1, \ldots, y_n)^t$ is a column vector of length n, $\mathbf{X}$ is the $n \times p$ matrix of explanatory variables, where the $[i, j]$ entry is the value of the j^{th} explanatory variable for y_i. Further, we define the parameters of our model to be $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^t$ the regression coefficients and the precision $\tau \in \mathbb{R}_+^*$. We make the following assumption on the conditional distribution of $\mathbf{y}$:

$$\mathbf{y}|\boldsymbol{\beta}, \tau \sim \text{MVN}\left(\mathbf{X}\boldsymbol{\beta}, \frac{1}{\tau}I\right).$$

This is the same as assuming that $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$, where the errors ϵ are iid normal random variables with mean 0 and variance $\frac{1}{\tau}$.

This is the linear regression setting, one of the simplest modeling methods in statistics which is appreciated for its simplicity and good performance in many situations. To fit a model to this setting we minimise the sum of squared errors $\sum(y - \hat{y})^2$, with $\hat{y} = \mathbf{X}\hat{\boldsymbol{\beta}}$. Simple matrix algebra then gives the Least Squares Estimate (LSE) for $\boldsymbol{\beta}$ to be: $\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$.

We can adapt this to a Bayesian setting quite simply with the introduction of priors, which allow us to have parameters with inherent errors, which in many physical situations is a more accurate representation of the world than classical statistics. We will begin by deriving the marginal distribution of $\mathbf{y}$ in terms of both priors on τ and $\boldsymbol{\beta}$.

Theorem II.1. *Given priors* $\boldsymbol{\beta}|\tau \sim Normal(\boldsymbol{\beta_0}, \frac{1}{\tau}\boldsymbol{\Sigma_0})$ *and* $\tau \sim Gamma(a_0, b_0)$, *the marginal likelihood of the responses under our Bayesian MLR model is:*

$$\mathbf{y} \sim \text{MVT}_{2a_0}\left(\mathbf{X}\boldsymbol{\beta_0}, \frac{b_0}{a_0}(I_n + \mathbf{X}\boldsymbol{\Sigma}_0\mathbf{X}^t)\right)$$

The difference with the classical setting is interesting. Note that we have a t distribution instead of a normal, whose degrees of freedom depend on the prior for the precision. This gives more flexibility to estimates in terms of the range of thickness of the tails, i.e. of the size of a confidence interval. Note that we also have a precision for $\mathbf{y}$ that depends on the mean of the prior for precision, and whose off-diagonal terms depend on the interaction between $\mathbf{X}$ and $\boldsymbol{\Sigma}_0$. This is a way to also relax the assumption of independence of errors by taking $\boldsymbol{\Sigma}_0$ to not be diagonal.

Now that we have an understanding of how in general terms $\mathbf{y}$ is represented in the Bayesian setting we will introduce a special case of Bayesian regression called Ridge regression which will allow us to better estimate $\boldsymbol{\beta}$.

iii. Ridge Regression

One of the major issues with MLR is the fact that it can only be performed if $\mathbf{X}^t\mathbf{X}$ is invertible. This leads to instability in estimates when $\mathbf{X}^t\mathbf{X}$ is near-singular, and in the worst case when $\mathbf{X}^t\mathbf{X}$ singular to uncomputable coefficients. Ridge Regression is one Bayesian method which circumvents this problem.

In Ridge Regression we choose the conditional prior of $\boldsymbol{\beta}$ given τ to have $\boldsymbol{\beta_0} = 0$ and $\boldsymbol{\Sigma_0} = \frac{I}{\tau\lambda}$, where λ is a fixed positive constant known as the shrinkage parameter. In effect, Ridge regression shrinks the β coefficients towards 0 as lambda increases and removes problems with singularity in $\mathbf{X}^t\mathbf{X}$. To see why we will examine the posteriors for $\boldsymbol{\beta}$ and τ.

Theorem II.2. *In Ridge regression as it has been defined above, we have:*

$$\boldsymbol{\beta}|\mathbf{y} \sim \text{MVT}\left(\boldsymbol{\beta_n}, \frac{b_n}{a_n}\boldsymbol{\Sigma_n}\right)$$

and

$$\tau|\mathbf{y} \sim \text{Gamma}(a_n, b_n)\,.$$

Where: $\boldsymbol{\beta_n} = (\mathbf{X}^t\mathbf{X} + I_p\lambda)^{-1}\mathbf{X}^t\mathbf{y}$, $\boldsymbol{\Sigma_n} = (\mathbf{X}^t\mathbf{X} + I_p\lambda)^{-1}$, $a_n = a_0 + \frac{n}{2}$, *and* $b_n = b_0 + \frac{1}{2}(\mathbf{y}^t\mathbf{y} - \boldsymbol{\beta}_n^t\boldsymbol{\Sigma}_n^{-1}\boldsymbol{\beta_n})$.

Notice that as lambda becomes large in the above theorem, the off-diagonal terms of $(\mathbf{X}^t\mathbf{X} + I_p\lambda)$ become negligible compared to the diagonal, so its inverse will be close to the null matrix, shrinking $\boldsymbol{\beta}$. Note also that the $I_p\lambda$ term ensures the above matrix is never singular, this solves our problem with matrix singularity. Later, we will apply Ridge regression to see its effectiveness in practice, but first we will outline the evaluation methods we will use throughout the analysis.

iv. Evaluation methods

One of the other advantages of a Bayesian regression approach is that it has a very intuitive prediction function called the posterior predictive:

$$f(y^*|\mathbf{y}) := \text{E}_{\boldsymbol{\theta}|\mathbf{y}}[f(y^*|\boldsymbol{\theta})] = \int f(y^*|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$$

This can help us find computationally effective method for estimates and errors using a closed form expression for the distribution of $\mathbf{y}^*|\mathbf{y}$. We derive this for Ridge regression in the following theorem.

Theorem II.3. *In Ridge regression, for k predictions* $\mathbf{y}^*$ *to be made with covariates* $\mathbf{X}^*$*:*

$$\mathbf{y}^*|\mathbf{y} \sim \text{MVT}_{2a_n}\left(\mathbf{X}^*\boldsymbol{\beta}_n, \frac{b_n}{a_n}(I_k + \mathbf{X}^*\boldsymbol{\Sigma}_n\mathbf{X}^{*t})\right)\,.$$

In practice, we have a_n large, so we will use the normal distribution. Of note is the fact that this can actually be used on values of $\mathbf{y}$ themselves. Doing so will help us gauge if the value y_i is an outlier based on the posterior p-value, i.e. $P(x > y_i)$. If this is abnormally small this suggests that y_i might be an outlier. The elegance of this method is that it is entirely self-contained and it doesn't need external computations like Cook's distance.

We now will introduce three evaluation methods: information criteria, evidence, and cross-validation.

There are two main information criteria, the Akaike[4] and the Bayesian[5] information criterion (AIC and BIC). Both are simple statistics that can be easily computed for a model to estimate how good of a fit it is. With log-likelihood $l(\theta|\mathfrak{M})$ we have: AIC $:= 2k - 2\max_\theta l(\theta)$ and BIC $:= k\log(n) - 2\max_\theta l(\theta)$. There has been lots of discussion of their properties, including for regression the study by Yang[6].

Bayesian approaches have another built in method which we will use for model selection: evidence. Looking back you might recall that Bayesian methods introduced $m(\mathbf{y})$ the marginal likelihood of $\mathbf{y}$. It is the integration constant $c = \int f(\mathbf{y}|\theta)\pi(\theta)d\theta$, which ensures the posterior is proper and it also acts as a measure of the quality of the model. To see how, let us rewrite $m(\mathbf{y}) := m(\mathbf{y}|\mathfrak{M})$, with $\mathfrak{M}$ denoting the model. Now applying Bayes theorem:

$$p(\theta|\mathbf{y},\mathfrak{M}) = \frac{p(\mathbf{y}|\theta,\mathfrak{M})p(\theta|\mathfrak{M})}{m(\mathbf{y}|\mathfrak{M})}$$

$$m(\mathbf{y}|\mathfrak{M}) = \frac{p(\mathbf{y}|\theta,\mathfrak{M})p(\theta|\mathfrak{M})}{p(\theta|\mathbf{y},\mathfrak{M})}$$

$$\log m(\mathbf{y}|\mathfrak{M}) = \log p(\mathbf{y}|\theta,\mathfrak{M}) + \log p(\theta|\mathfrak{M}) - \log p(\theta|\mathbf{y},\mathfrak{M})\,.$$

Note that the first term is a measure of fit, with higher values being better, while the

other two are penalty terms which make the evidence decrease as p the number of parameters increases. Therefore the evidence acts similarly to information criteria, and incorporates both goodness of fit and simplicity.

We now introduce a completely different process, cross-validation. Cross-validation is not a statistic that allows model comparison as the above are. It is a general heuristic to avoid the requirement for testing or validation sets, when the data is rare or expensive. The key concept of cross-validation is to exploit repeated random sampling of the dataset with parallel training. It is particularly useful for calibrating learning parameters $\hat{\alpha}$, and estimating error in a model f. In this paper we will use the standard 5-fold cross-validation, where we divide the dataset into five slices with random sampling. Then we train f on four folds and evaluate the results on the last one, and repeat in parallel so each fold is tested upon. This then gives a good measure of the effect of the tuning of $\hat{\alpha}$ or out-of-sample error of f.

While AIC and BIC can be applied to any MLR model, evidence is only possible in the Bayesian regression setting. Both give us a good idea of the quality of the model as an explanation, which we will use throughout our following analysis. Cross-validation however, will allow us to evaluate a model based on its predictive power. Which method is preferable depends on the problem at hand. In this analysis we will present them all, starting with evidence, followed later by the AIC/BIC model and then by selection using minimisation of the cross-validated Mean Squared Error (MSE).

III. Analysis

i. Maximal Evidence Model

In the Bayesian setting, which allows for the evidence, we will proceed via ridge regression. First we must determine the optimal λ. Note that the value of λ we consider optimal is the one that maximises the evidence. Since we require a model to determine the evidence we will use the full model, and assume that the model with maximum evidence is maximal regardless of the value of λ.

Using a method called Empirical Bayes, we will begin by optimising the shrinkage parameter in terms of the evidence using numerical optimisation. We find $\lambda_{\text{opt}} \simeq 0.1475$, which we will carry implicitly from now on, and which gives the full model a log-evidence of -2446.105. Figure 1 presents the underlying curve, with the evidence plotted on a log scale. One clearly sees that the evidence is higher at λ_{opt} than at 0, the "ridge-less" Bayesian regression, which simplifies to classical regression. To evaluate the difference in predictive power, we will, using the same folds compute the MSE for both values of λ, and include them in Table 1. We note that we have seen large improvements in the MSE by using a ridge regression. If we want to better our predictive ability we could optimise the value of λ with respect to the cross-validated MSE, obtaining $\lambda^* = 0.1525$. Doing so results in a negligible decrease in evidence (-0.009) and a negligible increase in predictive power, as seen in Table 2. The proximity of these two optimal shrinkages suggests evidence as a very good measure of both model fit and predictive power.

A further interesting development would be to write a variant algorithm which does stepwise selection using the evidence as the criteria. We will do so after presenting models from stepwise selection using the usual AIC and BIC.

ii. Stepwise Selected Models

Stepwise selection is a simple frequentist model selection method. It is customary to do it backwards, starting with the full model. Then, at each step of the algorithm we either remove from or add back to the model which ever variable leads to the greatest improvement in a criterion. When the AIC or BIC is used this leads to a non-greedy algorithm which converges to a minimum which is a model that is both simple and a good fit. At hand the key difference is going to be which criterion to choose. In this problem, use of the AIC leads to much more complex models as can be seen in Table 3. Note that `X.2` denotes a squared variable and `Y.Z` denotes the interaction of `Y` and `Z`.

In order to compare these with the previ-

ous model, let's compute the evidence we would find for a ridge regression using the variables for each of these two models, and their actual cross-validated MSE. We include the results in Table 4, which as the reader can see are underwhelming. Both the evidence and predictive power are much lower for both of these models compared to the full Ridge regression. Stepwise selection can sometimes generate worse predictions than a full model, as it is only really effective when there is a small subset of the parameters which account for most of the predictive power of the full model.

Now, let us compare these classical methods with our proprietary algorithm, which one can find in the final Appendix D. It's output is a model containing 13 parameters: `Int`, `sex`, `bmi`, `ltg`, `map`, `hdl`, `bmi.map`, `age.sex`, `glu`, `bmi.2`, `glu.2`, `tch`, `age.map`, with an evidence of -2435.786. A small increase but which is more significant than AIC and BIC. In terms of out-of-sample MSE, using cross-validation we come to an estimated MSE of 2825.106, with $\sigma = 378.25$. This is very competitive to BIC, the best performing of our models so far in terms of MSE as seen in Table 4.

iii. LASSO and LAR

To further better our predictions we can combine shrinkage, and variable selection in order to benefit from the best of both worlds. In this subsection we will provide two connected methods to do this, the LASSO and Least Angle Regression (LAR). In order to understand the LASSO we need to review Ridge regression in terms of the formulation which is usually done for Ordinary Least Squares (OLS) regression. See that in OLS we wish to find $\boldsymbol{\beta}$ which minimises the training MSE: $\frac{1}{n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^t(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$. The difference in Bayesian regression and LASSO is the existence of a regularisation term, which shrinks the estimates. In Bayesian regression we minimise $\frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2$, with $\lambda\|\boldsymbol{\beta}\|_2$ the regularisation term. See that when minimising this we recover $\boldsymbol{\beta}_n = (\mathbf{X}^t\mathbf{X} + I_p\lambda)^{-1}\mathbf{X}^t\mathbf{y}$ as we did in Theorem II.2 with the full distributions. The simplest presentation of the LASSO comes by giving the target it minimises, $\frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \kappa\|\boldsymbol{\beta}\|_1$[1] where κ depends on the data and a constant t. The significance of κ becomes evident when we rewrite this in a slightly different form using t: we minimise $\frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$, with the constraint that $\|\boldsymbol{\beta}\|_1 \leq t$. These two equations allow us to both connect the LASSO with previous methods, and to understand what it effectively does. To understand why the LASSO pushes some of the coefficients to be zero, we present Figure 2. One can see that the dashed contours of the two norms are different but why this difference is so dramatic is maybe not so evident. See that at the corners of the L_1 norm, one $\boldsymbol{\beta}$ is zero. Note how the red contour's first contact with the blue dotted line is at a corner for the L_1 norm but not for the L_2. this means that when minimising the surface this contour represents jointly with the L_1 norm we will get a null beta while in the L_2 norm we will get two non-zero betas. The "prominence" of the corners under L_1 norm is thus what leads to the much more frequent zero coefficients in the LASSO compared to Ridge regression. Intuitively, it is simply much more likely that the solution exists at a one of the four points with a zero β under the L_1 norm!

Theory aside, we will not apply the LASSO to this dataset and instead opt for LAR, which is another closely related method with better computational efficiency (equal to OLS[3]). LAR, like LASSO, combines variable selection with shrinkage, but has other benefits[2],[3].

LAR, again like LASSO doesn't output a single model, but rather a space of models dependent on the value of a parameter. To choose the best model we will use cross-validation of the MSE to select the step at which to terminate the LAR selection algorithm, as per *The ESL*[3] where more details on LAR and LASSO can be found. The cross-validation plot in Figure 3 shows that we should select the 20$^{\text{th}}$ step in the algorithm which gives us the model in Table 5. To compare the model obtained by LAR with the previous ones we can not proceed as above, as we can't in good faith compute the evidence for it as it uses a fundamentally different regularisation process. Whereas we can obtain

[1] Recall that $\|x\|_1 := \sum_i |x_i|$.

the OLS from Ridge regression with $\lambda = 0$, so using $\lambda_{\text{opt}} \simeq 0$ and computing the evidence is not shocking, this does not hold for LAR. Thus we compare them solely on the base of the MSE which is visible to be minimal at 20 around 3100, which is competitive with the result from full Ridge regression, but comes slightly short of outperforming AIC and BIC.

IV. Diagnostics

To finish our analysis we perform diagnostics and outlier detection, first using the predictive as outlined in section II.iv. Obtaining the posterior p-value for all points we find 18 points with a p-value less that 0.05, when we would expect to find 22 (Table 6). Out of these points, none seem to be notable outliers, either in the response nor in the covariates, thus they are accepted into the fold of this study. Using Cook's Distance on a full linear model found 8 points with $D \geq 4/n$ (Table 7). None of these were found to be abnormal, and thus they have not been excluded.

V. Conclusion

The best predictive model out of all the ones tried here is the stepwise selection model using BIC, with only 2% behind the result of our algorithm, which is the model with highest evidence. We have seen throughout that while evidence and predictive power aren't equivalent, they do seem heavily correlated. However from a practical standpoint, the LASSO and LAR are, for the purpose of computer based statistical linear modeling quite superior. They can be optimised for prediction using cross-validation, build on all the benefits of Ridge regression with both stability in estimates, robustness to correlated covariates, and are bundled into a process which is as fast as OLS regression. While Ridge regression works well and produces both a good MSE and has the value of the intrinsic evidence measure on this problem, in any instance where prediction is valuable the possibility of direct optimisation with cross-validated MSE in LAR makes it the most desirable method, as rewriting tailored search algorithms like our own is unfeasible.

We have noted that Bayesian approaches are competitive with classical ones, and that evidence seems to be a good measure of both model specification and prediction, providing incentive for the use of Bayesian regression over classical regression. But classical methods are not to be discarded so easily as the BIC and LASSO have shown.

A further interesting development[7] would be to attempt a similar comparison on Ensemble methods, for instance comparing the use of evidence or AIC and BIC as weights for the vote of several models obtained in this analysis, or otherwise. This method is more common in a classification problem so we might want to shift to a logistic regression setting instead of a linear one. In any case we would like to have a small number of models with differences in predictors, so as to reduce their independent bias and through averaging or voting to reduce their overall variance. This might be a complicated set-up as we would need a situation where there is high evidence for several notably different models without one model being vastly better at predictions than any other as this would remove the need for Ensemble methods. This could perhaps happen if the loss surface in the space of possible models were multi-modal or had a large area of near-zero gradient around its minimum, allowing for either several near minimal models spread across different local minima, or for several different models with loss reasonably near the minimum. A study of the actual conditions required to make such Ensemble methods worthwhile would also be interesting.

A. Tables

Fold	λ_{opt}	$\lambda = 0$	Difference
1	2806.80	3861.70	-1054.91
2	2941.78	4249.43	-1307.66
3	3503.98	4027.97	-524.00
4	3280.50	3508.66	-228.16
5	2661.66	3306.33	-644.68
Mean	3038.94	3790.82	-751.88

Table 1: *Comparison of MSE for different values of λ.*

Fold	λ^*	$\lambda = 0$	Difference (λ^* & 0)	λ_{opt}	Difference (λ^* & λ_{opt})
1	2807.75	3861.70	-1053.95	2806.80	0.95
2	2941.66	4249.43	-1307.78	2941.77	-0.11
3	3502.91	4027.97	-525.06	3503.98	-1.06
4	3279.03	3508.66	-229.63	3280.50	-1.47
5	2663.10	3306.33	-643.24	2661.66	1.44
Mean	3038.89	3790.82	-751.93	3038.94	-0.05

Table 2: *Comparison of MSE for different values of λ. (2)*

AIC	
Covariate	Coefficient
Int	152.13
sex	-266.35
bmi	496.08
map	343.02
tc	-857.70
ldl	683.59
ltg	972.75
age.2	83.51
tc.2	6187.14
ldl.2	3933.56
hdl.2	1053.37
ltg.2	1488.01
glu.2	195.50
age.sex	181.20
age.tc	-101.89
age.hdl	110.54
age.ltg	182.88
sex.map	84.15
bmi.map	171.33
map.glu	-119.87
tc.ldl	-9647.61
tc.hdl	-2947.87
tc.ltg	-4486.51
ldl.hdl	2330.72
ldl.ltg	3458.90
hdl.ltg	1568.96

BIC	
Covariate	Coefficient
Int	152.13
sex	-243.26
bmi	498.45
map	323.51
tc	-720.08
ldl	529.12
ltg	929.02
ltg.2	497.58
glu.2	133.19
age.sex	212.09
bmi.map	140.83
tc.ltg	-729.91
ldl.ltg	615.15
hdl.ltg	269.76

Table 3: *Models obtained by Stepwise selection on the* AIC *and* BIC*.*

Source	log Evidence	CV-MSE	σ(MSE)
Full Ridge Regression	-2446.105	3038.94	346.51
Stepwise with AIC	-2453.18	2831.13	433.08
Stepwise with BIC	-2449.684	2763.16	393.81

Table 4: *Using the previously obtained value of λ comparing evidence and MSE for all three models.*

Covariate	Coefficient
Int	152.03
sex	-143.49
bmi	499.52
map	269.72
hdl	-207.56
ltg	472.28
glu	30.59
age.2	23.42
bmi.2	45.30
glu.2	81.28
age.sex	120.20
age.map	30.17
age.ltg	16.54
age.glu	8.27
sex.map	16.64
bmi.map	96.98
map.hdl	10.01
tc.tch	-4.50

Table 5: *LAR model for step* 20*, the minimal MSE.*

Line	Posterior p-value
10	0.0134
38	0.0053
103	0.0041
151	0.0381
153	0.0360
191	0.0171
223	0.0136
233	0.0276
281	0.0322
283	0.0465
291	0.0309
305	0.0153
360	0.0083
361	0.0262
365	0.0439
379	0.0403
396	0.0306
405	0.0210

Table 6: *Points with extreme values of the posterior p-value.*

Line	D
2	0.7355
12	0.0207
24	0.0481
36	0.0106
59	0.0340
83	0.0149
261	0.0251
323	0.0096

Table 7: *Outliers in Cook's Distance.*

B. Figures

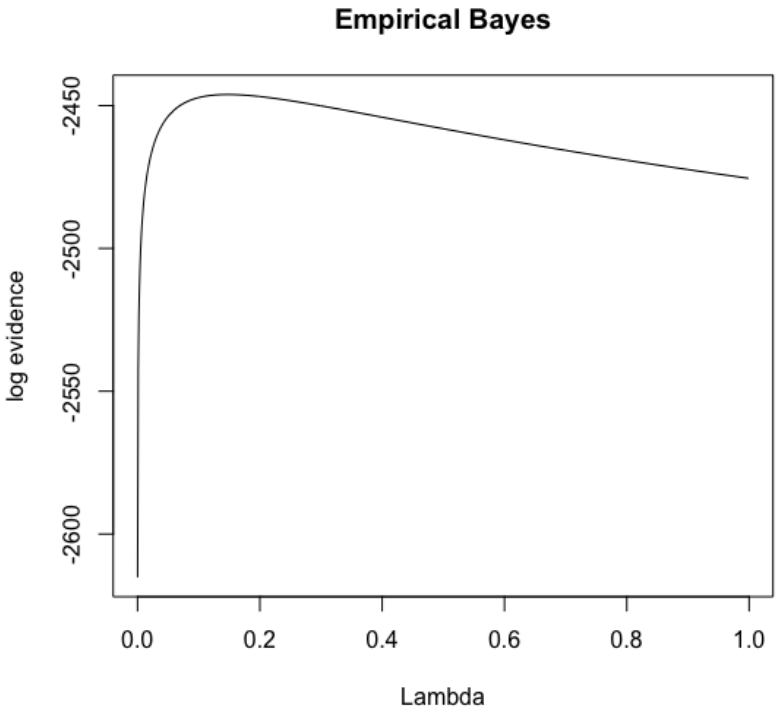

Figure 1: *Log-evidence for a range of values of λ.*

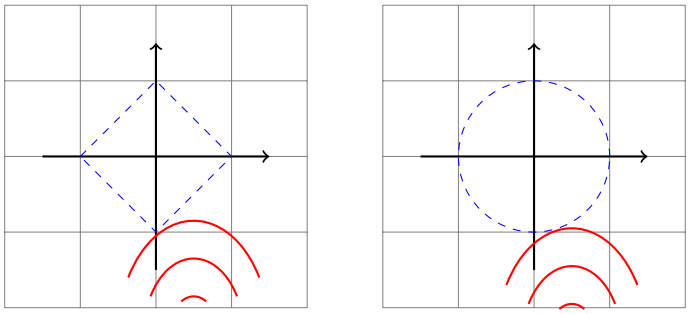

Figure 2: *Interaction between a contour (red) and the L_1 norm (left) and L_2 norm (right) boundaries for* 1.

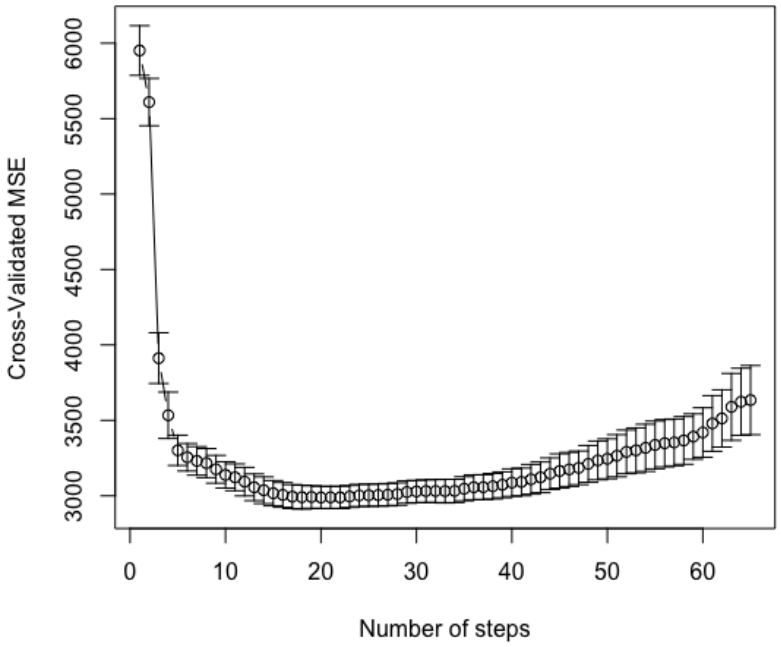

Figure 3: *C-V MSE for each step of the LAR algorithm.*

C. Proofs

Proof of Theorem II.1:

Lemma 1. *Given priors* $\boldsymbol{\beta}|\tau \sim Normal(\boldsymbol{\beta_0}, \frac{1}{\tau}\boldsymbol{\Sigma_0})$ *and* $\tau \sim Gamma(a_0, b_0)$*, under bayesian MLR, we have:*

$$\mathbf{y}|\tau \sim \text{MVN}\left(\mathbf{X}\boldsymbol{\beta_0}, \frac{1}{\tau}(I + \mathbf{X}\boldsymbol{\Sigma_0}\mathbf{X})\right).$$

Proof of lemma 1. Note that by the definition of our model, we can write that given τ, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}_1$ where $\boldsymbol{\epsilon}_1 \sim \text{MVN}(0, \frac{1}{\tau}I)$. Further, still with τ given, we have $\boldsymbol{\beta} = \boldsymbol{\beta_0} + \boldsymbol{\epsilon}_2$, where $\boldsymbol{\epsilon}_2 \sim \text{MVN}(0, \frac{1}{\tau}\boldsymbol{\Sigma_0})$. Now note that by the closure property of normal distributions, we know immediately that $\mathbf{y} = \mathbf{X}\boldsymbol{\beta_0} + \mathbf{X}\boldsymbol{\epsilon}_2 + \boldsymbol{\epsilon}_1$ will be normal with mean $\mathbf{X}\boldsymbol{\beta_0}$. Further, $\text{Var}(\mathbf{y}) = \mathbf{X}\,\text{Var}(\boldsymbol{\epsilon}_2)\,\mathbf{X}^t + \frac{1}{\tau}I = \frac{1}{\tau}(I + \mathbf{X}\boldsymbol{\Sigma_0}\mathbf{X}^t)$. □

To continue our proof, we will apply the law of total probability to the pdf from lemma 1:

$$\begin{aligned}
f(\mathbf{y}) &= \int_0^\infty f(\mathbf{y}|\tau)\pi(\tau)d\tau \\
&= \int_0^\infty \sqrt{\det\left(2\pi\frac{I + \mathbf{X}\boldsymbol{\Sigma_0}\mathbf{X}^t}{\tau}\right)}^{-1} \exp\left\{-\frac{\tau}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta_0})^t(I + \mathbf{X}\boldsymbol{\Sigma_0}\mathbf{X}^t)^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta_0})\right\}\pi(\tau)d\tau \\
&= \sqrt{2^n\pi^n\det(I + \mathbf{X}\boldsymbol{\Sigma_0}\mathbf{X}^t)}^{-1} \times \frac{b_0^{a_0}}{\Gamma(a_0)}\int_0^\infty \tau^{a_0 + \frac{n}{2} - 1}\exp\left\{-\tau\left(b_0 + \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta_0})^t(I + \mathbf{X}\boldsymbol{\Sigma_0}\mathbf{X}^t)^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta_0})\right)\right\}d\tau\,.
\end{aligned}$$

Recognising the integral as a gamma integral we obtain:

$$
\begin{aligned}
f(\mathbf{y}) &= \sqrt{2^n \pi^n \det(I + \mathbf{X}\boldsymbol{\Sigma_0}\mathbf{X}^t)}^{-1} \times \frac{b_0^{a_0} \times \Gamma(\frac{2a_0+n}{2} - 1)}{\Gamma(a_0)} \times \left(\frac{1}{b_0 + \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta_0})^t (I + \mathbf{X}\boldsymbol{\Sigma_0}\mathbf{X}^t)^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta_0})} \right)^{\frac{2a_0+n}{2}} \\
&= \frac{\Gamma(\frac{2a_0+n}{2})}{(2a_0)^{\frac{n}{2}} \pi^{\frac{n}{2}} \det(I + \mathbf{X}\boldsymbol{\Sigma_0}\mathbf{X}^t)^{\frac{1}{2}} \Gamma(a_0)} \times a_0^{\frac{n}{2}} b_0^{a_0 - a_0 - n/2} \times \left(1 + \frac{1}{2a_0} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta_0})^t \frac{a_0}{b_0} (I + \mathbf{X}\boldsymbol{\Sigma_0}\mathbf{X}^t)^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta_0}) \right)^{-\frac{2a_0+n}{2}} \\
&= \frac{\Gamma(\frac{2a_0+n}{2})}{(2a_0)^{\frac{n}{2}} \pi^{\frac{n}{2}} \det\left(\frac{b_0}{a_0}(I + \mathbf{X}\boldsymbol{\Sigma_0}\mathbf{X}^t)\right)^{\frac{1}{2}} \Gamma(a_0)} \left(1 + \frac{1}{2a_0} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta_0})^t \left(\frac{b_0}{a_0}(I + \mathbf{X}\boldsymbol{\Sigma_0}\mathbf{X}^t) \right)^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta_0}) \right)^{-\frac{2a_0+n}{2}}.
\end{aligned}
$$

Thus we can recognise the distribution of $\mathbf{y}$ as that of a multivariate t distribution with mean $\mathbf{X}\boldsymbol{\beta_0}$, variance $\frac{b_0}{a_0}(I + \mathbf{X}\boldsymbol{\Sigma_0}\mathbf{X}^t)$, and $2a_0$ degrees of freedom. □

Proof of Theorem II.2: By Bayes' Theorem we can obtain an expression for the posteriors:

$$
\begin{aligned}
\pi(\boldsymbol{\beta}, \tau | \mathbf{y}) &= \frac{f(\mathbf{y}|\boldsymbol{\beta}, \tau)\pi(\boldsymbol{\beta}, \tau)}{m(\mathbf{y})} \\
&\propto f(\mathbf{y}|\boldsymbol{\beta}, \tau)\pi(\boldsymbol{\beta}|\tau)\pi(\tau) \\
&\propto \tau^{\frac{n}{2}} \exp\left\{ -\frac{\tau}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^t (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\} \sqrt{\det\left(\frac{I_p}{\tau\lambda}\right)}^{-1} \exp\left\{ -\frac{\tau}{2}\boldsymbol{\beta}^t I_p \lambda \boldsymbol{\beta} \right\} \tau^{a_0 - 1} \exp\left\{ -b_0 \tau \right\} \\
&\propto \tau^{\frac{n+p}{2} + a_0 - 1} \exp\left\{ -\tau \left(\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^t (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\beta}^t I_p \lambda \boldsymbol{\beta}}{2} + b_0 \right) \right\}.
\end{aligned}
$$

Integrating out tau

$$
\begin{aligned}
\pi(\boldsymbol{\beta}|\mathbf{y}) &\propto \int_0^\infty \tau^{\frac{n+p}{2} + a_0 - 1} \exp\left\{ -\tau \left(\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^t (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\beta}^t I_p \lambda \boldsymbol{\beta}}{2} + b_0 \right) \right\} d\tau \\
&\propto \left(\frac{1}{2} \left((\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^t (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\beta}^t I_p \lambda \boldsymbol{\beta} \right) + b_0 \right)^{-\frac{2a_n+p}{2}} \\
&\propto \left(b_0 + \frac{1}{2} \left(\mathbf{y}^t\mathbf{y} + \boldsymbol{\beta}^t \mathbf{X}^t \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\beta}^t I_p \lambda \boldsymbol{\beta} - \boldsymbol{\beta}^t \mathbf{X}^t \mathbf{y} - \mathbf{y}^t \mathbf{X} \boldsymbol{\beta} \right) \right)^{-\frac{2a_n+p}{2}} \\
&\propto \left(b_0 + \frac{1}{2} \left(\mathbf{y}^t\mathbf{y} + \boldsymbol{\beta}^t (\mathbf{X}^t \mathbf{X} + I_p \lambda) \boldsymbol{\beta} - \mathbf{y}^t \mathbf{X} \boldsymbol{\beta} - \boldsymbol{\beta}^t \mathbf{X}^t \mathbf{y} \right) \right)^{-\frac{2a_n+p}{2}} \\
&\propto \left(b_0 + \frac{1}{2}\mathbf{y}^t\mathbf{y} + \frac{1}{2} \left(\boldsymbol{\beta}^t (\mathbf{X}^t \mathbf{X} + I_p \lambda) \boldsymbol{\beta} - \mathbf{y}^t \mathbf{X} \boldsymbol{\beta} - \boldsymbol{\beta}^t \mathbf{X}^t \mathbf{y} \right) \right)^{-\frac{2a_n+p}{2}} \\
&\propto \left(b_0 + \frac{1}{2}\mathbf{y}^t\mathbf{y} + \frac{1}{2} \left(\boldsymbol{\beta}^t \boldsymbol{\Sigma}_n^{-1} \boldsymbol{\beta} - \boldsymbol{\beta}_n^t \boldsymbol{\Sigma}_n^{-1} \boldsymbol{\beta} - \boldsymbol{\beta}^t \boldsymbol{\Sigma}_n^{-1} \boldsymbol{\beta}_n \right) \right)^{-\frac{2a_n+p}{2}} \\
&\propto \left(b_n + \frac{1}{2} \left(\boldsymbol{\beta}^t \boldsymbol{\Sigma}_n^{-1} \boldsymbol{\beta} - \boldsymbol{\beta}_n^t \boldsymbol{\Sigma}_n^{-1} \boldsymbol{\beta} - \boldsymbol{\beta}^t \boldsymbol{\Sigma}_n^{-1} \boldsymbol{\beta}_n + \boldsymbol{\beta}_n^t \boldsymbol{\Sigma}_n^{-1} \boldsymbol{\beta}_n \right) \right)^{-\frac{2a_n+p}{2}} \\
&\propto \left(b_n + \frac{1}{2} (\boldsymbol{\beta} - \boldsymbol{\beta}_n)^t \boldsymbol{\Sigma}_n^{-1} (\boldsymbol{\beta} - \boldsymbol{\beta}_n) \right)^{-\frac{2a_n+p}{2}} \\
&\propto \left(1 + \frac{1}{2a_n} (\boldsymbol{\beta} - \boldsymbol{\beta}_n)^t \left(\frac{a_n}{b_n} \boldsymbol{\Sigma}_n^{-1} \right) (\boldsymbol{\beta} - \boldsymbol{\beta}_n) \right)^{-\frac{2a_n+p}{2}}.
\end{aligned}
$$

As this is the kernel of a multivariate t density, this shows that:

$$\boldsymbol{\beta}|\mathbf{y} \sim \text{MVT}_{2a_n}\left(\boldsymbol{\beta}_n, \frac{b_n}{a_n}\boldsymbol{\Sigma}_n\right)$$

Now Turning to τ, we make use of lemma 1, and noting that $\boldsymbol{\beta}|\tau \sim \text{N}\left(0, \frac{I_p}{\tau\lambda}\right)$ here, of theorem II.1

$$\begin{aligned}
\pi(\tau|\mathbf{y}) &= \int \pi(\boldsymbol{\beta}, \tau|\mathbf{y}) d\boldsymbol{\beta} \\
&\propto \int \tau^{\frac{n+p}{2}+a_0-1} \exp\left\{-\tau\left(\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^t(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\beta}^t I_p \lambda \boldsymbol{\beta}}{2} + b_0\right)\right\} d\boldsymbol{\beta} \\
&\propto \int \tau^{\frac{n+p}{2}+a_0-1} \exp\{-\tau b_0\} \exp\left\{-\frac{\tau}{2}\left(\mathbf{y}^t\mathbf{y} - \mathbf{y}^t\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^t\mathbf{X}^t\mathbf{y} + \boldsymbol{\beta}^t\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^t I_p\lambda\boldsymbol{\beta}\right)\right\} d\boldsymbol{\beta}\,.
\end{aligned}$$

We rewrite the second exponential using the definitions of $\boldsymbol{\beta}_n$ and $\boldsymbol{\Sigma}_n$, like we did above for $\boldsymbol{\beta}$:

$$-\mathbf{y}^t\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^t\mathbf{X}^t\mathbf{y} + \boldsymbol{\beta}^t\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^t I_p\lambda\boldsymbol{\beta} = -\boldsymbol{\beta}_n^t\boldsymbol{\Sigma}_n^{-1}\boldsymbol{\beta} - \boldsymbol{\beta}^t\boldsymbol{\Sigma}_n^{-1}\boldsymbol{\beta}_n + \boldsymbol{\beta}\boldsymbol{\Sigma}_n^{-1}\boldsymbol{\beta}\,.$$

We now have, adding and subtracting $\boldsymbol{\beta}_n^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_n$:

$$\pi(\tau|\mathbf{y}) \propto \int \tau^{\frac{n+p}{2}+a_0-1} \exp\left\{-\tau\left(b_0 + \frac{1}{2}(\mathbf{y}^t\mathbf{y} - \boldsymbol{\beta}_n^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_n)\right)\right\} \exp\left\{-\frac{\tau}{2}\left((\boldsymbol{\beta}-\boldsymbol{\beta}_n)^t\boldsymbol{\Sigma}_n^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_n)\right)\right\} d\boldsymbol{\beta}\,.$$

Integrating out $\boldsymbol{\beta}$ as the integral of a multivariate normal with mean $\boldsymbol{\beta}_n$ and covariance $\frac{1}{\tau}\boldsymbol{\Sigma}_n$:

$$\begin{aligned}
\pi(\tau|\mathbf{y}) &\propto \tau^{\frac{n+p}{2}+a_0-1} \exp\left\{-\tau\left(b_0 + \frac{1}{2}(\mathbf{y}^t\mathbf{y} - \boldsymbol{\beta}_n^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_n\right)\right\} \sqrt{\det\left(2\pi\frac{1}{\tau}\boldsymbol{\Sigma}_n\right)} \\
&\propto \tau^{\frac{n+p}{2}+a_0-1-\frac{p}{2}} \exp\left\{-\tau\left(b_0 + \frac{1}{2}(\mathbf{y}^t\mathbf{y} - \boldsymbol{\beta}_n^t\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_n\right)\right\} \\
&\propto \tau^{a_n-1}\exp(-b_n\tau)\,.
\end{aligned}$$

This is recognisable as the kernel of a Gamma density, thus we have shown that $\tau|\mathbf{y} \sim \text{Gamma}\,(a_n, b_n)$. □

D. Code

```
ev.alg=function(DATA,y){

  #initiation steps
  all.names=as.vector(names(DATA)) #
  names.curr=c("Int","sex") #initialise first name set with just Int and sex
  lm.curr=lm(y~.-1,data=DATA[,names.curr]) #init LM
  ev.curr=Ev.model(l.max$maximum,a=2,b=2,names.curr,data=DATA)
  k=c()
  e=1 #initial epsilon
  while (e>10^(-3)){ #until convergence
    cand.names= setdiff(all.names,names.curr) # candidate names
    ev.curr.store=ev.curr #store
    names.curr.store=names.curr #store
    j=1 #counter
    for (n in cand.names){ #forward step
      ev.curr=Ev.model(l.max$maximum,a=2,b=2,names.curr,data=DATA) #ev of model
      next.names=as.vector(c(names.curr.store,n)) #the names of tried model
      ev.next=Ev.model(lambda = l.max$maximum, names=next.names, data=DATA)
                #evidence of the tried model
      k[j]=ev.next #store
      j=j+1
      if (ev.next >= ev.curr){ #trivial list maximisation
        ev.curr=ev.next
        names.curr=next.names
      }
    }
     for (m in names.curr){#remove terms
      ev.curr=Ev.model(l.max$maximum,a=2,b=2,names.curr,data=DATA)
      next.names=setdiff(names.curr,m)
      ev.next=Ev.model(lambda = l.max$maximum, names=next.names, data=DATA)
      }
      k[j]=ev.next
      j=j+1
      if (ev.next >= ev.curr){
        ev.curr=ev.next
        names.curr=next.names
        }
    }
    e=ev.curr-ev.curr.store #epsilon
  }
 return(list(names=names.curr,evidence=ev.curr))
}
```

Figure 4: *R implementation for the evidence-stepwise selection algorithm.*

References

[1] Hocking, R. R. (1976). *A Biometrics Invited Paper. The Analysis and Selection of Variables in Linear Regression.*, Biometrics, 32(1), 1-49.

[2] Efron, B., Hastie, T., Johnstone,I., and Tibshirani, R. (2004). *Least angle regression*, The Annals of statistics 32.2: 407-499.

[3] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*, Springer Series in Statistics. Springer New Tork Inc., New York.

[4] Akaike, H. (1974). *A new look at the statistical model identification*, IEEE Transactions on Automatic Control, 19 (6): 716?723.

[5] Schwartz, G.E. (1978). *Estimating the dimension of a model*, Annals of Statistics, 6 (2): 461?464.

[6] Yang, Y. (2005). *Can the strengths of AIC and BIC be shared?*, Biometrika, 92: 937?950.

[7] Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). *Bayesian model averaging for linear regression models.* Journal of the American Statistical Association, 92(437), 179-191.