May 2023

Comparative Study of Variable Selection Methods for Genetic Data

Anna-Lena Kubillus
University of Wisconsin-Milwaukee

COMPARATIVE STUDY OF VARIABLE SELECTION METHODS FOR GENETIC DATA

by

Anna-Lena Kubillus

A Thesis Submitted in
Partial Fulfillment of the
Requirements for the Degree of

Master of Science
in Mathematics

at
The University of Wisconsin-Milwaukee
May 2023

ABSTRACT

COMPARATIVE STUDY OF VARIABLE SELECTION METHODS FOR
GENETIC DATA

by

Anna-Lena Kubillus

The University of Wisconsin-Milwaukee, 2023
Under the Supervision of Professor David Spade

Association studies for genetic data are essential to understand the genetic basis of complex traits. However, analyzing such high-dimensional data needs suitable feature selection methods. For this reason, we compare three methods, Lasso Regression, Bayesian Lasso Regression, and Ridge Regression combined with significance tests, to identify the most effective method for modeling quantitative trait expression in genetic data. All methods are applied to both simulated and real genetic data and evaluated in terms of various measures of model performance, such as the mean absolute error, the mean squared error, the Akaike information criterion, and the Bayesian information criterion. The results show that all methods perform better than the ordinary least squares model on the prediction of future data. Moreover, the Lasso Regression outperforms all methods in terms of execution time and simplicity of the model, which therefore leads to better interpretability and makes it the best choice for association studies. Overall this thesis provides valuable insights into the strength and limitations of existing feature selection methods for modeling quantitative trait expression and highlights its importance in association studies for genetic data.

TABLE OF CONTENTS

LIST OF FIGURES

1 Introduction

1.1 Motivation and Goal

Association studies with SNPs have become a powerful tool in genetic research for identifying genetic variants associated with complex diseases and traits, as seen for example in [26]. Single nucleotide polymorphisms, or SNPs, are the most common type of genetic variation and represent a single nucleotide change in the DNA sequence of an individual's genome. These variations can have significant effects on gene expression and ultimately disease susceptibility [9].

Association studies compare the frequency of a particular SNP between two groups of individuals. One group has a specific disease or trait and the other hasn't. If the frequency of the SNP is significantly different between the two groups, it indicates that the SNP may be associated with the disease or trait [26]. This association can then be used to identify potential genetic targets for drug development, disease prevention, and personalized medicine [22].

An appropriate feature selection method is particularly essential in the processing and analysis of genetic data due to the large number of SNPs included. This should help to provide better interpretability and reduce the variance of the trained models, which is often caused by overfitting the model to the training data.

In this thesis, we compare the performance of three feature selection methods for effective analysis of genetic data and modeling quantitative trait expression. Specifically, we consider Lasso regression, Bayesian Lasso regression, and a combination of Ridge regression and significance testing. All of these methods are expected to reduce the variance of the model and ensure better interpretability compared to the ordinary least squares model. We aim to identify the most suitable method to develop a relationship between the SNPs and the observed trait value, such as blood pressure. By doing so, we can contribute to the development of new methods for effective analysis of genetic data and a better understanding of the genetic basis of complex traits.

1.2 Structure

Afterward, chapter two describes the data sets that form the base of the thesis and are available for training and evaluation of the models. Furthermore, the third chapter describes the theoretical basics of the different models that are compared as well as the measures used for the comparison. The main part of the thesis deals with implementing the models in R in chapter four and the subsequent simulation studies, a comparison of the models, and an evalution of the results in chapter five. Finally, the results of the thesis are summarized in the conclusion.

2 Dataset

The identification of genetic variants, in the context of this thesis as single nucleotide polymorphisms (SNPs), which are associated with different traits of an individual, is a common field of research that has been performed very successfully in many cases (for example in [3][13][1][21]). The aim here is to use a suitable method to find a relationship between the SNPs and the observed characteristic, such as blood pressure. Thereby it is possible to draw conclusions about possible cardiovascular diseases based on the genetic information of an individual [25]. This fact should serve as a starting point of this thesis. The aim is to use a suitable feature selection method to infer from the originally numerous SNPs occurring in the DNA strand to a few central ones that are responsible for the expression of the trait. The data sets of this thesis comprise an original data set of bred mice as well as synthetically produced data, which are both randomly divided into 80% training and 20% test data for building and validation of the models.

2.1 Original Data

The real data set was obtained from a genome wide assoication study in which data from 288 bred NMRI mice were collected [28]. Genomic DNA was isolated from tail biopsies by phenol-chloroform extraction and 581,672 SNP genotypes were determined per mouse. The exact isolation and preparation process can be read here [27]. After removing and merging identical SNPs, 44428 unique SNP genotypes ultimately remained, which are now used as features. At 8 weeks of age, blood pressure was also measured using a tail cuff. There were 100 measurement cycles, in which the systolic and diastolic blood pressure and the mean arterial pressure were measured. Thereupon, all outliers with a standard deviation > 2 were removed from the mean of the measurements and the final average of the remaining values was assigned as the result for each mouse. For the further analysis, we will now restrict the thesis to the prediction of systolic blood pressure using the available SNP genotypes.

2.2 Synthetic generated data

The synthetically generated data has a similar structure. For this, 1000 unphased genotype sequences of length 50,000 with minor allele frequency between 10% and 30% are simulated using the *ms* function without selection in R. After all constant features are removed, a locus is randomly selected to generate quantitative feature data. Similar to what was described by Thompson and Kubatko in [23], these data are then generated along the evolutionary tree at the selected SNP sites. Here, based on major (A) and minor (a) alleles occurring in SNP sites, a distinction is made between 3 different combinations of these: AA coded as 0, the heterogeneous allele pairs Aa and aA is coded as 1, and aa is coded as 2. Thus, a higher mean trait value suggests a higher proportion of minor alleles. If, in addition, q denotes the probability of such a minor allele and the Hardy-Weinberg equilibrium is assumed, the following probabilities result: AA occurs with probability $(1-q)^2$, heterogeneous pairs with $2q(1-q)$, and aa with q^2. The process assumes an additive model and therefore uses the following generalized Hansen model, which is a generalized version of the Ornstein-Uhlenbeck process:

$$dY_i(t) = \alpha(\Theta_i(t) - Y_I(t)) + \sigma_Y dB_i(t) \qquad \text{where } \Theta_i(t) = \begin{cases} \Theta_0, & ifS_i(t) = 0 \\ \Theta_1, & ifS_i(t) = 1 \\ \Theta_2, & ifS_i(t) = 2 \end{cases} . \tag{1}$$

$Y_i(t)$ represents the trait value for lineage i at time t. Θ is the mean trait value and is set to $\Theta_1 = 80$, $\Theta_2 = 100$ and $\Theta_3 = 120$ in the simulation. $\alpha = 5$ represents the strength of selection towards Θ and $\sigma_Y = 10$ is the standard deviation of the model per unit time. $B_i(t)$ is a Brownian motion process for line i, making the values $dB_i(t)$ independent identically normally distributed random variables with mean 0 and standard deviation dt for a small time interval dt. It should be noted that the correlation between SNP sites resulting from the generation of genetic data in this way is not considered in this model. After performing the methods just described datasets with 1000 genotype sequences and around 25000 to 30000 SNP features each are left for further consideration.

3 Theory

3.1 Feature Selection Methods

Data sets in which the target variable has an underlying normal distribution are initially very well suited for the prediction with a linear regression model. However, especially with high dimensional data, one often encounters two problems in particular. First, models that use numerous variables to predict a target variable are very difficult to interpret. Removing variables that are redundant or receive no new information makes the model simpler without losing information [11]. In addition, prediction accuracy plays a critical role. In general, the higher the complexity of a model, the lower the squared bias, but the higher the variance of these models [16]. Due to this so-called bias-variance tradeoff, the prediction accuracy of ordinary least squares estimates is low, especially for high-dimensional data. To avoid overfitting, the original model is adjusted by a regularization term, which reduces the variance significantly and can partially perform variable selection [19].

The SNP data set is also affected by this problem. Presumably, not all SNP variants influence the blood pressure of an individual. Nevertheless, they are all included in the ordinary least squares model, which on the one hand makes it very difficult to interpret the model and to attribute much, little, or no influence to certain variants. On the other hand, the danger of overfitting is also very high. If more than 40,000 predictors are used to train a model from a data set with only about 300 data points, the model will adapt so strongly to the data points due to the enormous imbalance that it will perform much worse later on future data. Therefore, three methods for feature selection and regularization are described below, which are applied and finally compared in this thesis.

3.1.1 Lasso

The Lasso (least absolute shrinkage and selection operator) technique first introduced by Tibshirani in 1996 [24] combines the aspect of feature selection with regularization. We start with an ordinary regression situation with data (x_i, y_i), $i = 1, ..., N$, where x_i represents the predictor variables and y_i the response. The OLS estimate is defined as:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_j \beta_j x_{ij} \right)^2 \right\} \tag{2}$$

Lasso regression now adds another constraint, called a penalty term, which is an upper bound (t) on the sum over all coefficients. This results in the following optimization problem for the determination of the Lasso estimates:

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_j \beta_j x_{ij} \right)^2 \right\} \qquad \text{subject to} \sum_j |\beta_j| \leq t[24]. \tag{3}$$

Moreover, it can be rewritten as follows:

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_j \beta_j x_{ij} \right)^2 + \lambda \sum_j |\beta_j| \right\} [19]. \tag{4}$$

The penalty term leads to a shrinkage of the coefficients, whereby the size of the λ controls the amount of shrinkage. Compared to other regularization methods, Lasso shrinks some coefficients to zero, especially in highly correlated groups, and these can then be excluded from the model. This improves the interpretability of the model and reduces the variability of the estimates [19].

3.1.2 Bayesian Lasso

An extension to the Lasso method is the Bayesian Lasso regression. Here, a different prior distribution of the β coefficients is chosen, which affects their posterior distribution. Tibshirani describes the posterior distribution of the β coefficients in the regular Lasso regression as

$$\hat{\beta}^{lasso} = \arg\max_{\beta} p(\beta|y, \sigma^2, \tau)[24] \tag{5}$$

The prior distribution of β_i is an independent Laplace distribution and the likelihood component follows a multivariate normal distribution:

$$\left.\begin{aligned} p(\beta|\tau) &= \left(\frac{\tau}{2}\right)^p exp(-\tau||\beta||_1) \\ p(y|\beta, \sigma^2) &= N(y|X\beta, \sigma^2 I_n) \end{aligned}\right.[15] \tag{6}$$

The version of the Bayesian Lasso regression model presented by Chris Hans in 2009 now specifies a scaled version of the double exponential distribution as a prior which leads to the following setting:

$$\left.\begin{aligned} p(\beta|\tau, \sigma^2) &= \left(\frac{\tau}{2\sigma}\right)^p exp(\tau\sigma^{-2}||\beta||_1) \\ p(y|\beta, \sigma^2, \tau) &= N(y|X\beta, \sigma^2 I_n) \end{aligned}\right.[15] \tag{7}$$

$N(y|X\beta, \sigma^2 I_n)$ fits a multivariate normal distribution with mean $X\beta$ and variance $\sigma^2 I_n$ evaluated at y. In addition, independent prior distributions $\sigma^{-2} \sim Ga(a, b)$ and $\tau \sim Ga(r, s)$ are specified for σ and τ, respectively, and thus the mode $p(\beta|\tau, \sigma^2)$ is the Lasso estimate with penalty parameter $\lambda = 2\tau\sigma$ [15].

From these assumptions, the resulting posterior distribution is obtained:

$$p(\beta_j|\beta_{-j}, \sigma^2, \tau, y) = p(y|\beta_j, \sigma^2, \tau)p(\beta_j|\sigma^2, \tau)p(\sigma^2)p(\tau) \tag{8}$$

Finally, observations of the respective distribution are drawn for β, σ^2 and τ using Gibbs sampling [15] or a reversible jump Markov Chain Monte Carlo algorithm.

3.1.3 Ridge regression with Significance Tests

Ridge regression is another approach that uses a different regularization term to reduce the variance of the OLS model. In contrast to Lasso regression, the L_2 norm is now used in the penalty term. Ridge regression thus minimizes the following quantity:

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j} \beta_j x_{ij} \right)^2 \right\} \qquad \text{subject to} \sum_{j} \beta_j^2 \leq t. \tag{9}$$

Which is equivalent to:

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j} \beta_j x_{ij} \right)^2 + \lambda \sum_{j} \beta_j^2 \right\}. [24] \tag{10}$$

A major disadvantage of this method is that in Ridge regression coefficients are shrunk and regularized, but due to the property of the L_2 norm rarely reach 0. This does not ensure better interpretability of models trained with high-dimensional data. In Figure 1 the minimization problem for the selection of the coefficients in Lasso and Ridge regression in the two-dimensional space is graphically represented. In the center of each ellipse is the coefficient β estimated by the OLS model. The constraint region, defined by the L_1 and L_2 norms, respectively, can be seen as a circle and rotated square. Considering the first point at which this touches the ellipses, it is quite possible in the Lasso regression that this occurs at one of the corners of the square. In the circular constraint of the Ridge regression, on the other hand, this happens extremely rarely without corners, which is why both coefficients are only shrunk but are still non-zero. [24]

Thus, no variables are excluded from the model by Ridge regression. To take advantage of the regularization and the associated reduction in the variability of the predictions in the model, this thesis also uses significance tests for the individual coefficients in addition to Ridge regression for feature selection. These tests should help to determine the importance of the individual predictors so that unimportant variables can be excluded from the model. Therefore the significance test presented by Cule, Vineis, and De Iorio in [8] is used, which is based on a t-test of individual coefficients and was specially constructed for high dimensional data. The following test statistic is defined with the help of the Ridge coefficients b_j and its standard error (se($\hat{\beta}_j^{\lambda}$)):

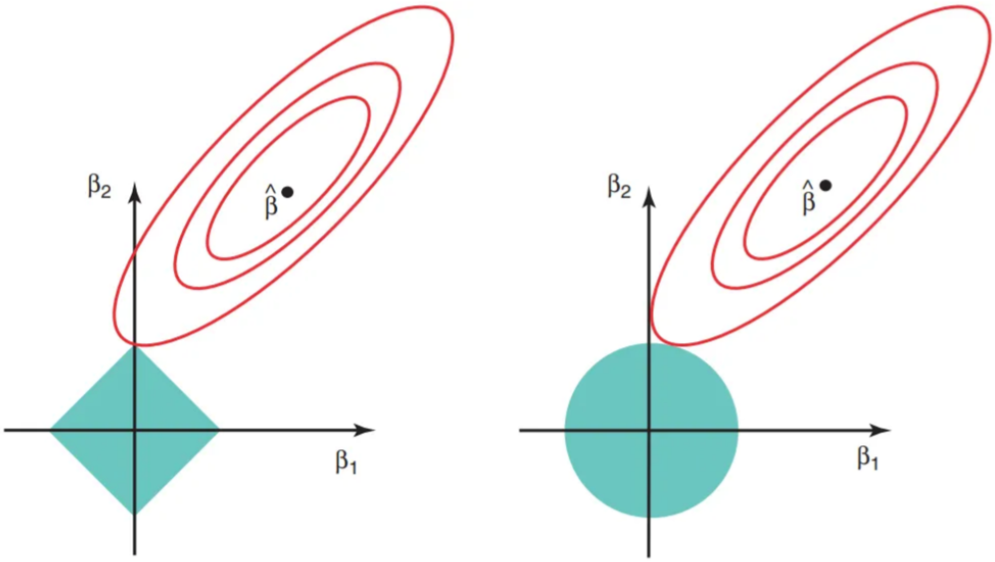

Figure 1: Lasso vs. Ridge - shrinkage process
[17]

$$T_\lambda = \frac{\hat{\beta}_j^\lambda}{se(\hat{\beta}_j^\lambda)} \tag{11}$$

The standard error is calculated using the root of the respective diagonal entry in the covariance matrix:

$$Var(\hat{\beta}^\lambda) = \sigma^2 (X^T X + \lambda I)^{-1} X^T X (X^T X + \lambda I)^{-1} \qquad \text{where } \hat{\sigma}^2 = \frac{(Y - X\hat{\beta})^T (Y - X\hat{\beta})}{\nu} \tag{12}$$

ν represents the residual effective degrees of freedom. In the original approach of Halawa and El Bassiouni [14] $\nu = n - m$ is used. Since for high dimensional data ν would be negative, the degrees of freedom are now obtained using the so-called ”hat matrix” $H = X(X^T X + \lambda I)^{-1} X^T$. Then

$$\nu = n - tr(H) \tag{13}$$

where $tr(H)$ is the trace of A.

For the significance test of a predictor, the null hypothesis H_0 , $T_\lambda \sim N(0,1)$ is assumed. [8] Under this assumption and a significance level of $\alpha = 0.01\%$, the test is then performed and non-significant variables are removed from the data set. The α was set that low to strongly reduce the large number of coefficients in the model to the most decisive SNP sites. Finally, a linear model is again trained and evaluated on the remaining predictors using Ridge regression.

3.2 Measures

To be able to examine the performance of the models described in the last subsection in terms of their predictive accuracy, the mean absolute error (MAE), root mean squared error (RMSE), Akaike information criterion (AIC) and Bayesian information criterion (BIC) are used.

3.2.1 RMSE and MAE

Root mean squared error (RMSE) and mean squared error (MAE) are both performance measures assessing the performance with the loss function which compares the predicted values to the actual values.

The RMSE here is the root of the mean squared error (MSE), which returns the mean squared difference of the predicted values to the actual values:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2[7]} \tag{14}$$

In contrast, the MAE calculates the mean of the absolute errors:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|[7] \tag{15}$$

In the end, both measure the prediction accuracy on future data and the lower each of these values is, the better the model.

3.2.2 AIC and BIC

In conjunction with the MAE and RMSE criteria, which calculate the predictive power of different models for future values based on predictions of test data, two further criteria will now be considered. These compare different models on the base of their likelihood function and thus assess, without using predictions of previously unknown test data, which model is most suitable in terms of interpretation, prediction, and subsequent use. One class of such alternative criteria is the penalized model selection criteria, which describe the goodness of a model essentially by two building terms. First, the maximum likelihood is considered, which reflects the performance concerning the known training data. However, this tends to favor larger models with many predictors. Especially with high dimensional data, a model tends to overfit the training data, which results in poor predictive power with respect to future data. Therefore, a second term is intended to penalize the increase of the log-likelihood by a higher number of predictors in the model. By describing two terms pulling in opposite directions, the tradeoff between model fit and model complexity is represented, which plays an important role regarding the prediction with genetic data [18].

Several such measures have been introduced in different literature, but this thesis is limited to the use of AIC and BIC.

AIC is the Akaike information criterion and is defined as follows:

$$AIC = -2ln(\hat{L}) + 2k, [2] \tag{16}$$

where k is the number of nonzero coefficients in the model and $ln(\hat{L})$ is the maximized value of the log-likelihood function of the model. If we compare different models, the model with a lower AIC value is preferred, since a higher maximum log-likelihood of the model leads to a lower AIC value. The second term counteracts this and penalizes large models that are difficult to interpret and have a strong tendency to overfit. The so-called Bayesian information criterion (BIC) pursues the same goal of describing the performance of a model using the maximum log-likelihood function, but its penalty term is somewhat different:

$$BIC = -2ln(\hat{L}) + k \cdot ln(n), [20] \tag{17}$$

where n is the number of data points in the training data set. In contrast to the AIC before, the BIC penalizes large models depending on the number of data points used for training and the number of predictors. Comparing the two penalty terms $2k$ and $ln(n)k$, $2k > ln(n)k$ is valid already from $n = \lceil e^2 \rceil = 8$ data points in the training data and thus the BIC conteracts overfitting even more.

4 Implementation in R

4.1 Lasso and Ridge regression

For the implementation of Lasso and Ridge regression in R, the common library `glmnet` is used because it has very efficient procedures to train linear, logistic, multinomial, possion and cox regression models [12] and has a significantly shorter runtime compared to other libraries. `Glmnet` solves the following optimization problem:

$$\min_{\beta_0,\beta}\frac{1}{N}\sum_{i=1}^{N} w_i l(y_i,\beta_0+\beta^T x_i)+\lambda[(1-\alpha)||\beta||_2^2/2+\alpha||\beta||_1] \tag{18}$$

Since this thesis is limited to the training of ordinary linear models, the negative log-likelihood contribution described by $l(y_i, \eta_i)$ for the i^{th} consideration refers to the Gaussian case $\frac{1}{2}(y_i - \beta_0 - \beta^T x_i)^2$. The regularization parameter λ is determined over a grid of values that covers the entire width of all possible solutions and from which the λ with the smallest error is determined. The larger λ, the greater the penalty for high coefficients in the model. With the help of the parameter nlambda, which was set to 500 in the implementation, the number of λ values available for selection is defined. Depending on which model is to be trained, the parameter α is set. It spans the bridge between Ridge regression (if $\alpha = 0$), where the regularization term including the L1 norm in 18 disappears and the whole weight lies on the first term, and the Lasso regression (if $\alpha = 1$), where only the L1 norm is considered[5]. Calling the `glmnet` function for Lasso regression is done by:

```
model <- glmnet(X, y,  alpha = 1, nlambda = 500, intercept = TRUE,
    standardize= FALSE)
```

If, on the other hand, a linear model is trained via Ridge regression, the following line is executed:

```
model <- glmnet(X, y,  alpha = 0, nlambda = 500, intercept = TRUE,
    standardize= FALSE)
```

In both cases, an ordinary linear regression model is fitted using the training data. The predict function can then be used to make predictions for the test data and determine the metrics used for evaluation:

```
 y_predicted <- predict(model , s = model$lambda[n_lambda] , newx = X_test)
```

4.2 Significance Tests

As described in chapter 3.1.3, the Ridge regression reduces the variance of the model and thus promises a better performance on future unknown data, but does not provide a better interpretability of the model. Therefore an additional procedure is needed. Significance tests for feature selection will now be used for this purpose. Due to the high number of features in both datasets, the computing effort is very high especially for calculations like $X^T \cdot X$ or for inverting matrices of an appropriate size, which is needed to create the ”hat”-matrix as well as in the calculation of the covariance matrix. Just for storing a double-precision matrix of size 45000x45000 around 16 GB RAM is needed. Chapter 3.1.3 outlines calculations that require storage of at least three such matrices. As a result, carrying out these calculations on a regular computer can prove to be prohibitively expensive or even unfeasible. To minimize this computational expenditure the matrices are converted into single precision matrices with the help of the `float` package [10].

```
X_single <- fl(X_double)
y_single <- fl(y_double)
```

This reduces both the time required for the calculations and the memory space. The example matrix from above now only requires around 8 GB. For inverting the matrices, the `chol2inv(A)` function from the matrix package is used, since this is more time-saving than the commonly used `solve()` function. For the significance test, the amount of the test value T_0 calculated from the β_i coefficient and its estimated standard deviation is compared with the 99.9% quantile of the t-distribution (t_{crit}) for each feature. If $|T_0| < t_{crit}$ holds, no significant linear relationship between the respective SNP genotypes and the systolic blood pressure can be detected and the feature is removed from the model. Finally, the model is trained and evaluated again with all remaining SNP genotypes.

4.3 Bayesian Lasso

Since the *glmnet* package does not support the Bayesian Lasso method, the `monovmn` package is used. This was excluded in the training for the other procedures due to higher runtime. With the help of the implemented blasso function, the model described in chapter 3.1.2 is trained on the training data:

```
 model <- blasso(X, y, T=10, icept=TRUE, normalize = FALSE, mprior=c(2000,
    38000))
```

The linear model here includes an intercept because the given data is not standardized. Due to computational expense, only 10 RJMCMC (T=10) samples from the Bayesian LASSO procedure are collected. In addition, a mprior distribution is specified. mprior is the prior for the number of regression coefficients that are not zero. Setting $mprior = c(2000, 38000)$ represents a Binomial distribution $Bin(m|n = M, p)$ where $p \sim Beta(2000, 38000)$. Therefore it sets the probability of inclusion for a coefficient to be on average 5%. [6] The regression coefficients of the model are then extracted and used to make predictions for the test data:

```
\label{predict_blasso}
      betas <- c(model$mu[t],model$beta[t,])
  X_with1 <- cbind(rep(1, nrow(X_test)), X_test)
  y_predicted <- X_with1%*%betas
```

Finally, the various measures of model performance are calculated, which are used in the following chapter to compare the various models.

5 Results

5.1 Simulation Studies

Using the measures described in 3.2 the performance in terms of predictive power is compared in this chapter. Therefore datasets with 1000 datapoints and usually between 25000 and 30000 predictors are generated and each of the models is trained 50 times. Due to computational expense for the significance tests, only 20000 predictors are picked randomly in each run.

model	time	lambda	number_coef	MAE	sqrt_MSE	AIC	BIC
ols	17.63	0.00	20001.00	10.45	13.35	-100990.59	-7280.95
lasso	7.41	0.10	218.68	8.12	10.20	-93479.98	-92459.92
ridge	1157.05	3988.41	19625.50	8.01	10.06	-29296.10	62654.27
blasso	358.63	215.07	552.60	8.34	10.53	3977.78	6567.77

Table 1: 'Mean value of the various measured variables from 50 simulations'

5.1.1 General Comparison

In table 1 the mean values for the respective measures, the number of predictors remaining in the final model including one additional for the intercept, the selected regularization parameter λ and the time required for one run from the 50 simulation rounds are presented. It is evident that Ridge regression and significance tests take a considerable amount of time for training and require a high computational effort, as discussed in chapter 4.2. Similarly the Bayesian Lasso trained by the `blasso` function of the `monovmn` package cannot keep up with the optimized algorithms of the `glmnet` package with regard to the time, demonstrated comparing the time per run for OLS and Lasso regression. The number of coefficients still contained in the final model is significantly reduced by both varieties of Lasso regression. While the Bayesian Lasso only retains about 550 parameters on average, the Lasso regression reduces the number of coefficients to just over 200. In both cases, the substantial reduction enhances the interpretability of the models, which is an essential goal of the feature selection.

On the other hand, the number of nonzero coefficients in the Ridge regression model remains almost the same even after significance tests with an uncommonly low significance level of $\alpha = 0.01$ are performed. On average, only just around 400 of the 20000 features considered were removed, indicating limited interpretability. This is likely due to the low correlation between individual predictors. A major advantage of Ridge regression is typically to detect highly correlated groups of features and to shrink all but one. Within the significance tests, these would be declared as not significant and thus removed from the model. However, the correlations between individual features in this study are generally very low, with only around 2.5% exhibiting a correlation of more than 0.9. This implies that the significance tests have minimal effect, and almost all coefficients remain in the model.

In the following chapters, each of the three feature selection methods is compared first with the ordinary least squares regression model, followed by comparisons among them.

5.1.2 Comparison Ordinary Least Squares versus different Feature Selection Methods

By examining both the MAE shown in Figure 2 and the RMSE shown in Figure 3, it is clear that all three models have improved. These figures compare the MAE and RMSE of the OLS model to those of the Lasso, Ridge, and Bayesian Lasso regression. For each simulation round, the bisector $y = x$ is plotted to indicate which model has the lower value. If a point is above the line, the MAE or RMSE of the OLS is lower, and if it is below, the opposite method performs better. Both MAE and RMSE are calculated using test data that the model has not previously seen, and therefore they measure the model's predictive accuracy on future data. The improvement is due to the fact that the OLS model is heavily overfitted, and this fit to the training data results in poorer predictive accuracy on the test data. Regularization, which is applied to Ridge regression as well as to Lasso and Bayesian Lasso procedures, helps improve the prediction of the test data, often at the expense of the model accuracy on the training data. While the OLS model is on average about 10.5 off

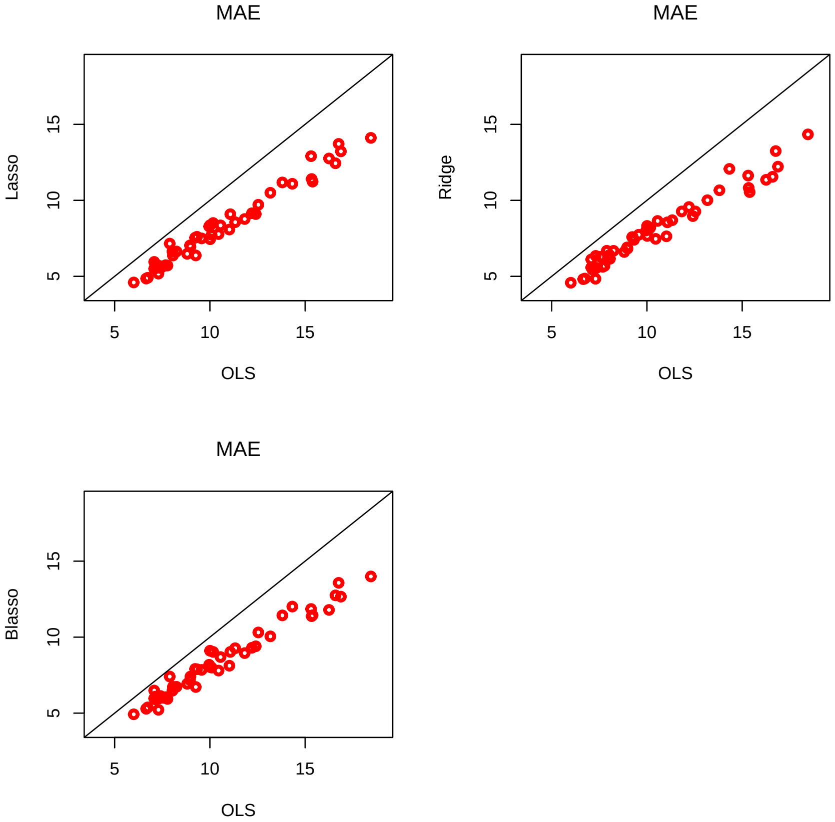

Figure 2: Scatterplots of MAE from OLS versus other Feature Selection methods. The panels present the following methods: (Top Left): Lasso, (Top Right): Ridge, (Bottom Left): Bayesian Lasso

the correct value, Lasso has a mean deviation of 8.1, Bayesian Lasso has 8.3, and Ridge regression has an even lower mean deviation of 8.0. Similar results can be observed for the RMSE, where Ridge regression shows an improvement from 13.4 to as low as 10.1. Both variants of the Lasso procedure have RMSE values around 10.3. Thus, the evaluation of these measures suggests a significant improvement of all models compared to the original OLS model.

If we also examine AIC and BIC, which evaluate the model purely based on maximum log-likelihood on the training data and the number of predictors included, we obtain slightly different results. Figures 4 and 5 compare AIC and BIC between the OLS model and Lasso, Ridge, and Bayesian Lasso regression. The left column again depicts the bisector to illustrate

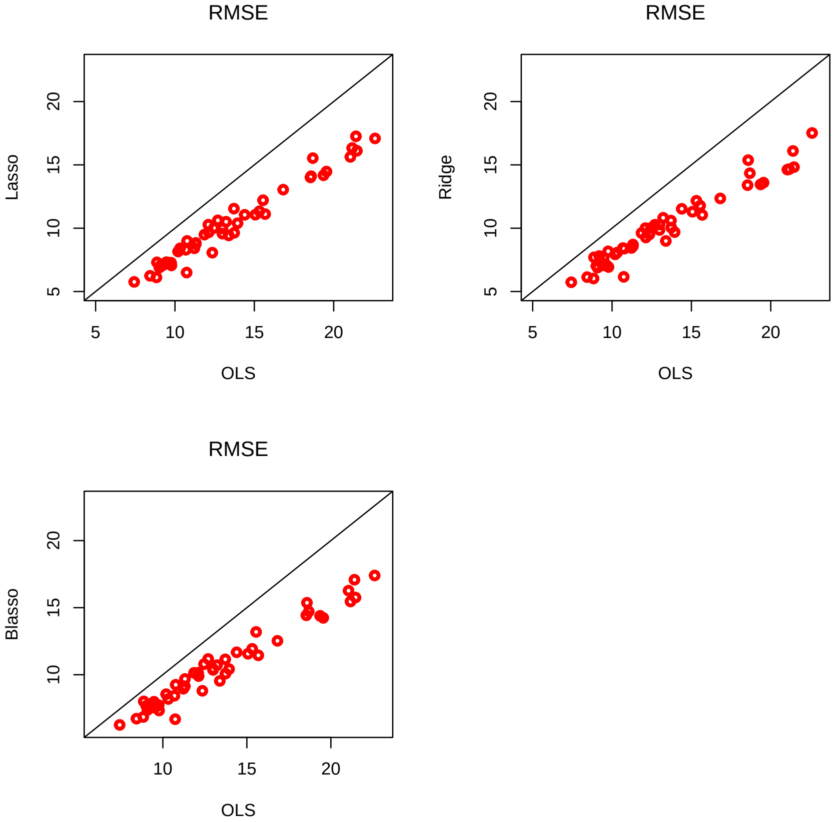

Figure 3: Scatterplots of RMSE from OLS versus other Feature Selection methods. The panels present the following methods: (Top Left): Lasso, (Top Right): Ridge, (Bottom Left): Bayesian Lasso

the direct comparison of the two with the same scaling of the axes, while the right column describes the exact distribution of the points on differently scaled axes.

Here, Ridge regression shows a clear result. In all 50 simulations, both the AIC and BIC for Ridge regression are higher than those of the OLS, which makes the OLS a better model regarding those measures. This is because strong overfitting occurs especially in the OLS model without regularization, resulting in a very high maximum log-likelihood on the training data. Thus, it minimizes both AIC and BIC more than it is the case with the Ridge regression with regularization to avoid overfitting. Additionally, both models have a similar number of non-zero coefficients, which makes the penalty term roughly the same. Consequently, Ridge regression performs worse than the OLS model in all 50 simulations.

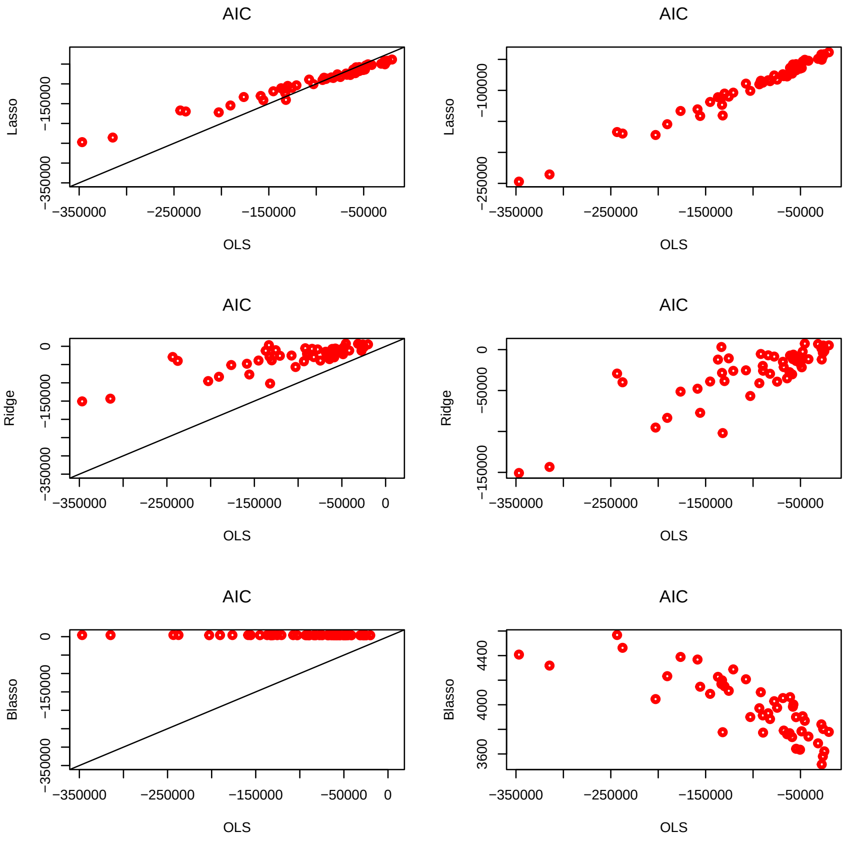

Figure 4: Scatterplots of AIC from OLS versus other Feature Selection methods. The panels present the following methods: (Top Row): Lasso, (Middle Row): Ridge, (Bottom Row): Bayesian Lasso

A completely different result is obtained when comparing the Lasso regression model and the OLS model. While the AIC still slightly tends towards the strongly overfitted OLS, the BIC prefers the Lasso model in almost all cases. The reason for this is the significantly different numbers of coefficients of both models. Looking first at the maximum log-likelihood, a very similar picture to the Ridge regression emerges between OLS and Lasso. The regularization term in the Lasso model counteracts overfitting and thus reduces the value of the maximum log-likelihood. In contrast, in Lasso regression, the penalty term for the number of parameters in the model is much smaller. In the AIC, this lower penalty is not as noticeable as it is in the BIC, which is why it tends to favor the OLS model. The BIC in Figure 5 reflects a clear advantage on the side of the Lasso model since in this measure, as

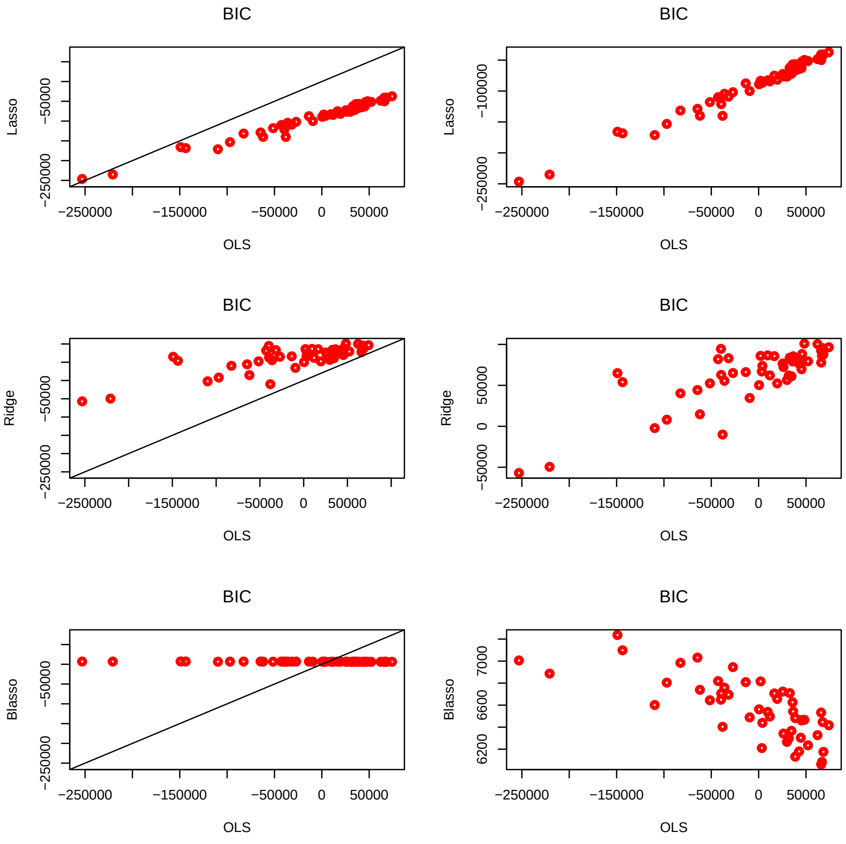

Figure 5: Scatterplots of BIC from OLS versus other Feature Selection methods. The panels present the following methods: (Top Row): Lasso, (Middle Row): Ridge, (Bottom Row): Bayesian Lasso

described earlier in Chapter 3.2.2, large models that tend to overfit the data are penalized stronger. Consequently, the significant advantage of the few coefficients in the Lasso model has a stronger impact.

Next, we analyze the comparison between OLS and Bayesian Lasso regression. The AIC results show a clear preference for OLS over Bayesian Lasso regression, which is consistent with the Ridge regression results. Interestingly, Bayesian Lasso is very stable in its performance across different simulated data sets. The left plots with equally scaled x- and y-axis show that all points seem to lie almost on a horizontal line. Only by adjusting the scaling of the axes, one can see some fluctuations in different runs, which are much smaller than in the OLS model. Looking at the BIC, the results shift in favor of the model with the smaller number of coefficients. There is no longer a clear tendency toward the OLS model, as there are runs with smaller BIC values in the Bayesian Lasso regression as well as in the OLS regression. Therefore, no clear winner emerges here.

In conclusion, the results confirm the assumption of overfitting in the OLS model. Looking only at the max log-likelihood of the models, the OLS performs best, and the other models can only keep up in a direct comparison if they are less penalized by their smaller number of coefficients in the model and achieve a similar value in the AIC or BIC. The regularization in the individual methods counteracts this overfitting and ensures better prediction accuracy of future unknown data than the OLS in all three feature selection methods.

5.1.3 Comparison Lasso versus Bayesian Lasso regression

In this section, we compare regular Lasso regression with Bayesian Lasso regression, as shown in Figure 6. Both the MAE and RMSE plots show a slight bias towards regular Lasso. In most runs, the Lasso regression model resulted in a comparatively smaller RMSE, reflected in the average value of 10.2 compared to 10.5 in Bayesian Lasso. The average MAE is also lower for regular Lasso, at 8.1 compared to 8.3 for Bayesian Lasso.

The comparison of AIC and BIC shows that the Lasso model outperforms the Bayesian Lasso in both cases. However, since both models were able to greatly reduce the number of predictors and now have a similar number of coefficients remaining, this also applies to the penalty term in both measures. The max log-likelihood of the Lasso model dominates, from which one can conclude a significantly better performance in terms of prediction accuracy on the training data. This may also be related to stronger regularization in the Bayesian Lasso model, explaining the significantly larger average regularization parameter (λ) in Bayesian Lasso of 215, compared to 0.1 in Lasso. As a result, the Bayesian Lasso model cannot adapt strongly enough to the training data and has a lower performance.

In conclusion, the Lasso regression model shows a better tendency than the Bayesian Lasso in all four measures, indicating that the Lasso makes better predictions based on

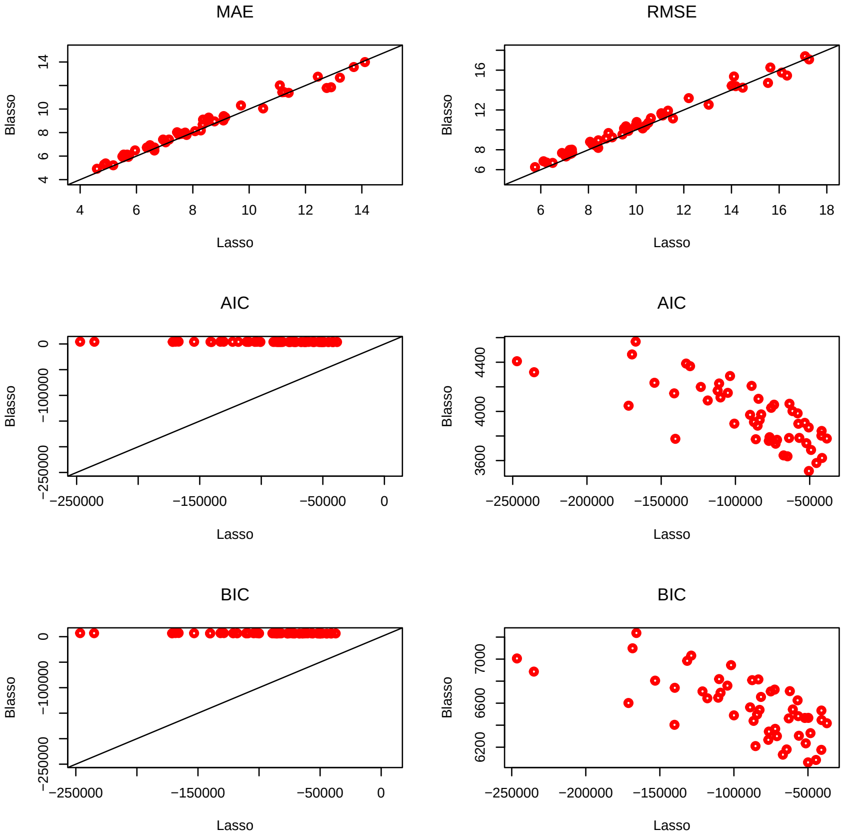

Figure 6: Scatterplots of comparison of different prediction measures between Lasso and Bayesian Lasso. The panels present the following methods: (Top Row): MAE and MSE, (Middle Row): AIC, (Bottom Row): BIC

this simulation study and therefore has a higher prediction accuracy. The Lasso model is also significantly faster to generate, averaging only 7.5 seconds compared to the Bayesian Lasso, which takes around 6 minutes per run. Finally, the Lasso regression retains only 220 coefficients on average, while the Bayesian Lasso retains around 550. Thus, regular Lasso guarantees even better interpretability and should be preferred to Bayesian Lasso based on the studies conducted here.

5.1.4 Comparison Lasso versus Ridge regression

In this section, we compare the performance of the Ridge regression model trained with significance tests with that of the Lasso model. The results of the simulation study presented

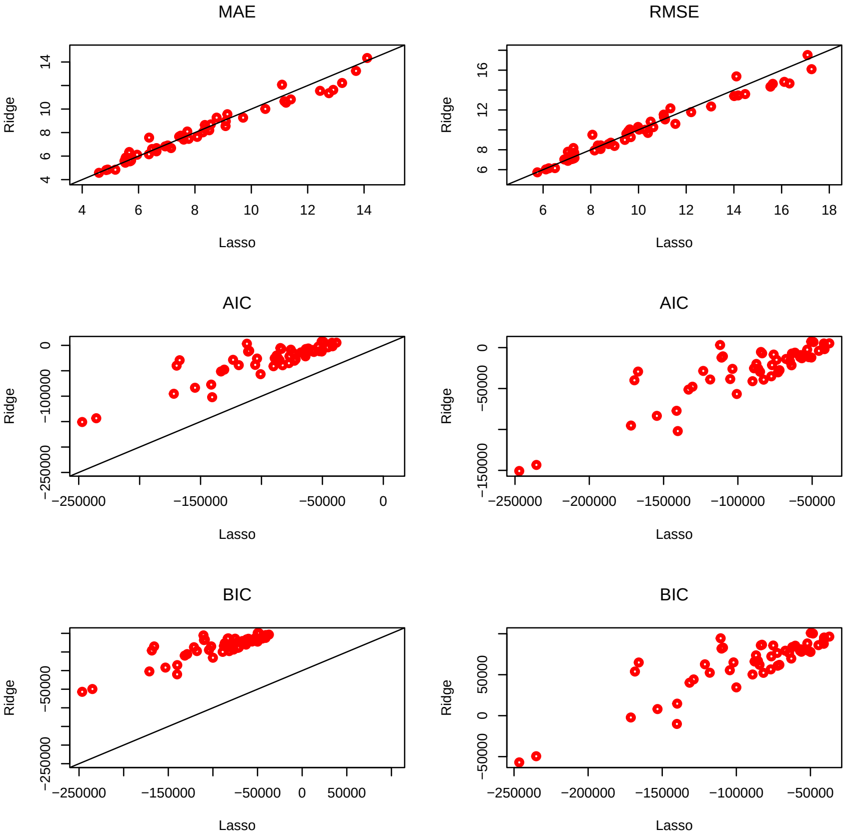

Figure 7: Scatterplots of comparison of different prediction measures between Lasso and Ridge regression. The panels present the following methods: (Top Row): MAE and MSE, (Middle Row): AIC, (Bottom Row): BIC

in Table 1 and Figure 7 indicate that both models perform similarly in terms of MAE and RMSE, with a slight preference for the Ridge regression model. However, the comparison based on AIC and BIC shows that the Lasso model outperforms the Ridge regression model. This can be attributed to the high number of remaining coefficients in the Ridge regression model, which leads to a much higher penalty term and subsequently results in a higher value of AIC and BIC. The high number of nonzero coefficients in the Ridge regression model also makes it less interpretable and difficult to filter out specific SNP features responsible for certain trait values.

In addition, the Lasso model has a significantly faster execution time, taking an average of only 7 seconds compared to the Ridge regression model which takes around 20 minutes due to the significance tests involved. Based on these results, we conclude that the Lasso regression model is the better model of the two and should be preferred over the Ridge regression model for its better performance, interpretability, and faster execution time.

5.2 Real Data Analysis

model	time	lambda	number_coef	MAE	sqrt_MSE	AIC	BIC
ols	0.07	0.00	20001	51.47	63.30	-1730215.63	-1660938.99
lasso	1.71	0.66	182	7.19	8.81	-1767733.74	-1767106.79
ridge	636.51	2323.48	19997	7.02	9.19	-1728471.01	-1659208.23
blasso	8.40	16.92	202	18.94	23.48	1219.40	1919.95

Table 2: 'Execution of the various measured variables using original data'

The simulation study results are now being applied to the original dataset to verify or refute them. Once again, a random selection of 20,000 out of the original 44,428 features was used to train the models due to limited RAM availability. Table 2 shows the results of the run. As previously observed, both MAE and RMSE improve significantly across all models compared to OLS. The Bayesian Lasso prediction reduce the deviation from over 50 to 19, and Ridge and Lasso regression to around 7. The same trend is observed for the RMSE.

However, the AIC and BIC shows somewhat different results. The Lasso regression model has the best values, with high prediction accuracy on the training data and only 182 predictors, leading to a low penalty term. The OLS model and Ridge regression follow closely behind, while the Bayesian Lasso model is very far behind, indicating poor prediction accuracy. The low values of the other models indicate a strong fit to the training data. All models are likely still impacted by overfitting. Due to the limited amount of 224 data points in the training dataset, it is hardly possible to train a model with such a high number of features without overfitting. Moreover, only 3 coefficients were removed from the Ridge regression, leading to poor interpretability.

Therefore, the results of the simulation study are confirmed, and Lasso regression is found to be the best model for the original data as well.

6 Conclusion

In this thesis, we compared three feature selection models to extend the ordinary least squares regression model. Based on the simulation study and real data analysis, the Lasso regression model emerged as the best, and we recommend it over the other models. In conjunction with Ridge regression, it provided the best results in terms of predictive power on future data, outperforming both the original OLS and Bayesian Lasso approaches. By using regularization, we were able to reduce the variance in the predictions and increase accuracy.

In addition, Lasso regression has proven its worth with regard to better interpretability of the model. While Ridge regression contained almost all coefficients even after performing significance tests, Lasso regression reduced the number of predictors to around 10% of the originally existing predictors. As mentioned earlier, besides prediction accuracy, it is important to be able to identify individual SNP features that can be used to draw conclusions about the level each feature is investgated in predicting the trait value. Lasso regression is better able to select SNPs that explain a high percentage of the variation in the trait values compared to Ridge regression and Bayesian Lasso, which produced more complex regression models in terms of the number of variables selected.

Another significant benefit of the Lasso approach is the time it takes to run. In the simulation studies, the Lasso model required almost 50 times less training time than the Bayesian Lasso and around 150 times less than Ridge regression with significance tests. It even finished in less than half the time of the OLS model.

Overall, Lasso regression is a good approach to filter out significant SNP features in genetic data. It provides relatively simple regression models with high prediction accuracy, explaining a large part of the variation in the trait values.

However, these models are strongly restricted to the given SNP features. When considering the blood pressure of a person, other factors such as diet or lifestyle also influence it in addition to genetic factors. Therefore, our regression models are currently very limited and would require additional information as covariates to make more precise, realistic statements and guarantee even more accurate results.

References

[1] International IBD Genetics Consortium (IIBDGC) et al. “Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis”. In: *Nature genetics* 45.11 (2013), pp. 1353–1360.

[2] Hirotugu Akaike. “A new look at the statistical model identification”. In: *IEEE transactions on automatic control* 19.6 (1974), pp. 716–723.

[3] David Altshuler, Mark J Daly, and Eric S Lander. “Genetic mapping in human disease”. In: *science* 322.5903 (2008), pp. 881–888.

[4] SangGyu An. *What are three approaches for variable selection and when to use which.* 2021. URL: https://medium.com/codex/what-are-three-approaches-for-variable-selection-and-when-to-use-which-54de12f32464.

[5] *An Introduction to glmnet.* URL: https://glmnet.stanford.edu/articles/glmnet.html (visited on 04/20/2023).

[6] *blasso: Bayesian Lasso/NG, Horseshoe, and Ridge Regression.* URL: https://www.rdocumentation.org/packages/monomvn/versions/1.9-17/topics/blasso (visited on 04/20/2023).

[7] Brad Boehmke and Brandon M Greenwell. *Hands-on machine learning with R.* CRC press, 2019.

[8] Erika Cule, Paolo Vineis, and Maria De Iorio. “Significance testing in ridge regression for genetic data”. In: *BMC bioinformatics* 12.1 (2011), pp. 1–15.

[9] Paul Elliott et al. “Genetic Loci associated with C-reactive protein levels and risk of coronary heart disease”. In: *Jama* 302.1 (2009), pp. 37–48.

[10] *float.* URL: https://cran.r-project.org/web/packages/float/readme/README.html (visited on 04/20/2023).

[11] Valeria Fonti and Eduard Belitser. “Feature selection using lasso”. In: *VU Amsterdam research paper in business analytics* 30 (2017), pp. 1–25.

[12] *glmnet-package: Elastic net model paths for some generalized linear models.* URL: https://www.rdocumentation.org/packages/glmnet/versions/4.1-6/topics/glmnet-package (visited on 04/20/2023).

[13] Hakon Hakonarson et al. “A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene”. In: *Nature* 448.7153 (2007), pp. 591–594.

[14] AM Halawa and MY El Bassiouni. “Tests of regression coefficients under ridge regression models”. In: *Journal of Statistical Computation and Simulation* 65.1-4 (2000), pp. 341–356.

[15] Chris Hans. “Bayesian lasso regression”. In: *Biometrika* 96.4 (2009), pp. 835–845.

[16] Trevor Hastie et al. *The elements of statistical learning: data mining, inference, and prediction.* Vol. 2. Springer, 2009, 27ff.

[17] Gareth James et al. *An introduction to statistical learning.* Vol. 112. Springer, 2013.

[18] Jouni Kuha. “AIC and BIC: Comparisons of assumptions and performance”. In: *Sociological methods & research* 33.2 (2004), pp. 188–229.

[19] R Muthukrishnan and R Rohini. “LASSO: A feature selection technique in predictive modeling for machine learning”. In: *2016 IEEE international conference on advances in computer applications (ICACA)*. IEEE. 2016, pp. 18–20.

[20] Andrew A Neath and Joseph E Cavanaugh. “The Bayesian information criterion: background, derivation, and applications”. In: *Wiley Interdisciplinary Reviews: Computational Statistics* 4.2 (2012), pp. 199–203.

[21] Stephan Ripke et al. “Genome-wide association analysis identifies 13 new risk loci for schizophrenia”. In: *Nature genetics* 45.10 (2013), pp. 1150–1159.

[22] Marylyn D Ritchie. “The success of pharmacogenomics in moving genetic association studies from bench to bedside: study design and implementation of precision medicine in the post-GWAS era”. In: *Human genetics* 131 (2012), pp. 1615–1626.

[23] Katherine L Thompson and Laura S Kubatko. “Using ancestral information to detect and localize quantitative trait loci in genome-wide association studies”. In: *BMC bioinformatics* 14.1 (2013), pp. 1–10.

[24] Robert Tibshirani. “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[25] William Valdar et al. “Genome-wide genetic association of complex traits in heterogeneous stock mice”. In: *Nature genetics* 38.8 (2006), pp. 879–887.

[26] Peter M. Visscher et al. “Five Years of GWAS Discovery”. In: *The American Journal of Human Genetics* 90.1 (2012), pp. 7–24. ISSN: 0002-9297. DOI: https://doi.org/10.1016/j.ajhg.2011.11.029. URL: https://www.sciencedirect.com/science/article/pii/S0002929711005337.

[27] Hyuna Yang et al. “A customized and versatile high-density genotyping array for the mouse”. In: *Nature methods* 6.9 (2009), pp. 663–666.

[28] Weidong Zhang et al. “Genome-wide association mapping of quantitative traits in outbred mice”. In: *G3: Genes— Genomes— Genetics* 2.2 (2012), pp. 167–174.

Appendix

Results of different simulation runs

	model	time_elapsed	lambda	number_coef	MAE	MSE	AIC	BIC
1	ols	14.616878	0.00000000	20001	6.001421	7.431362	-27552.796	66064.297
2	ols	13.916986	0.00000000	20001	9.321401	12.129847	-47674.674	45866.995
3	ols	13.786416	0.00000000	20001	7.077374	8.865100	-131997.736	-38057.051
4	ols	17.369682	0.00000000	20001	7.298584	10.735459	-27954.414	65986.271
5	ols	16.753671	0.00000000	20001	15.313889	18.684540	-132689.196	-39477.693
6	ols	18.201634	0.000000e+00	20001	16.761570	21.401887	-243361.428	-149053.760
7	ols	15.921513	0.000000e+00	20001	7.348833	9.453972	-41581.933	52135.286
8	ols	15.998817	0.000000e+00	20001	10.014019	12.989133	-145238.698	-51471.604
9	ols	10.373602	0.000000e+00	20001	9.224154	11.857573	-102901.693	-9461.034
10	ols	21.097041	0.000000e+00	20001	11.023246	13.942341	-60966.357	32875.318
11	ols	25.123200	0.000000e+00	20001	10.170953	12.698646	-89418.462	3379.055
12	ols	11.007853	0.000000e+00	20001	11.826441	15.065195	-125609.148	-31767.474
13	ols	16.172426	0.000000e+00	20001	6.655393	8.833073	-26378.138	66858.957
14	ols	13.729877	0.000000e+00	20001	7.142289	9.082018	-50470.175	42741.328
15	ols	10.852059	0.000000e+00	20001	7.680023	9.614322	-20107.317	74103.145
16	ols	11.858047	0.000000e+00	20001	10.559538	13.221779	-84036.707	9580.387
17	ols	31.136451	0.000000e+00	20001	9.258403	12.352847	-77682.006	16577.118
18	ols	16.996481	0.000000e+00	20001	11.328487	14.392582	-202784.111	-109726.850
19	ols	21.126842	0.000000e+00	20001	8.034842	10.322816	-25960.070	67606.772
20	ols	18.490323	0.000000e+00	20001	7.310539	9.193470	-31458.652	62183.520
21	ols	17.794987	0.000000e+00	20001	16.254732	21.167393	-158546.994	-64606.309
22	ols	18.454485	0.000000e+00	20001	16.878687	21.463701	-176416.621	-82574.946
23	ols	22.641029	0.000000e+00	20001	15.392264	19.359089	-133378.115	-39786.132
24	ols	18.797181	0.000000e+00	20001	11.072970	13.713888	-68500.924	25439.761
25	ols	16.896996	0.000000e+00	20001	7.406634	9.474694	-67619.440	26172.546
26	ols	14.920576	0.000000e+00	20001	8.979773	11.315630	-74588.170	19622.292
27	ols	20.375002	0.000000e+00	20001	15.342529	19.545697	-120853.596	-27136.377
28	ols	12.422207	0.000000e+00	20001	7.075277	9.014304	-64128.662	29910.545
29	ols	27.388570	0.000000e+00	20001	13.175283	16.818973	-107501.316	-13610.075
30	ols	16.347067	0.000000e+00	20001	8.806781	11.221452	-82085.390	11879.971
31	ols	14.591568	0.000000e+00	20001	7.891554	10.765409	-57447.974	36443.267
32	ols	9.544040	0.000000e+00	20001	14.329265	18.576929	-314603.299	-220736.825
33	ols	35.293637	0.000000e+00	20001	10.455914	13.390677	-54691.883	38698.079
34	ols	13.888631	0.000000e+00	20001	9.566026	13.730640	-89705.766	3986.4681
35	ols	17.192207	0.000000e+00	20001	12.409735	15.681776	-130104.103	-36138.7421
36	ols	20.992627	0.000000e+00	20001	8.246924	10.687019	-45421.504	48295.7154
37	ols	15.375957	0.000000e+00	20001	8.058221	10.197021	-61484.283	31905.6785
38	ols	17.735829	0.000000e+00	20001	12.543518	15.549852	-155759.415	-62017.2427
39	ols	14.296490	0.000000e+00	20001	10.006170	12.454998	-93362.560	454.2853
40	ols	20.442156	0.000000e+00	20001	12.207423	15.317373	-137002.136	-42791.6738
41	ols	11.892152	0.000000e+00	20001	10.088770	12.993076	-91810.595	2055.8785
42	ols	17.893412	0.000000e+00	20001	13.805192	18.539734	-190462.114	-96970.886
43	ols	23.649440	0.000000e+00	20001	18.452422	22.604854	-346863.589	-253096.495
44	ols	23.579784	0.000000e+00	20001	8.957317	11.314205	-48953.796	44512.164
45	ols	13.515563	0.000000e+00	20001	7.138155	9.402669	-58774.091	34615.871
46	ols	11.288497	0.000000e+00	20001	9.960601	12.103532	-57963.704	35853.141
47	ols	12.490184	0.000000e+00	20001	16.593652	21.050314	-237604.006	-143688.027
48	ols	16.384085	0.000000e+00	20001	7.772846	9.775943	-48629.584	45311.101
49	ols	11.343042	0.000000e+00	20001	7.560268	9.756015	-54599.187	38637.908
50	ols	39.403907	0.000000e+00	20001	6.738507	8.431422	-24842.941	68368.562

Table 3: 'Results of the various measured variables from 50 simulations of OLS Regression'

	model	time_elapsed	lambda	number_coef	MAE	MSE	AIC	BIC
1	lasso	7.567101	0.08960683	129	4.595769	5.758988	-50545.636	-49946.486
2	lasso	7.634124	0.07520341	265	7.602628	9.674539	-53331.498	-52096.748
3	lasso	7.308425	0.17620912	76	5.949244	7.305271	-140465.254	-140112.976
4	lasso	7.387386	0.07376484	217	5.175994	6.497758	-41942.321	-40927.761
5	lasso	7.677062	0.06582020	454	12.906985	15.532396	-123315.318	-121204.078
6	lasso	7.779537	1.224248e-01	283	13.714201	17.255921	-167176.144	-165846.406
7	lasso	7.389230	9.255598e-02	163	5.686924	7.316505	-52244.109	-51485.000
8	lasso	7.274265	1.512184e-01	152	7.441339	9.571571	-118635.362	-117927.421
9	lasso	7.045667	1.185646e-01	160	7.540223	9.493561	-100733.423	-99990.570
10	lasso	7.665625	6.855255e-02	293	8.079199	10.394451	-63611.754	-62241.665
11	lasso	10.345891	1.167136e-01	187	8.508495	10.618970	-86281.673	-85418.656
12	lasso	7.505492	9.844764e-02	229	8.769698	11.065776	-110076.351	-109006.556
13	lasso	7.563363	7.440821e-02	191	4.852300	6.116468	-45466.068	-44580.316
14	lasso	7.493786	1.210391e-01	121	5.529645	7.036740	-64953.258	-64393.989
15	lasso	7.826938	6.502580e-02	253	5.735721	7.281906	-38460.058	-37273.006
16	lasso	7.623199	1.017930e-01	209	8.352104	10.510594	-83663.607	-82689.989
17	lasso	8.041437	1.083880e-01	169	6.374483	8.068052	-75901.366	-75109.590
18	lasso	7.415827	1.744942e-01	161	8.563829	11.057948	-171932.658	-171188.200
19	lasso	7.817608	5.844348e-02	284	6.636179	8.409690	-42113.421	-40789.450
20	lasso	7.616690	8.231429e-02	164	5.670216	7.039596	-49071.685	-48308.501
21	lasso	8.074948	7.956352e-02	373	12.753738	16.331155	-130551.645	-128804.348
22	lasso	8.619729	9.759105e-02	356	13.221284	16.113993	-133212.222	-131546.533
23	lasso	7.699366	8.598361e-02	294	11.231602	14.177090	-111953.606	-110582.484
24	lasso	7.742377	6.534785e-02	317	9.086569	11.545577	-73969.849	-72485.586
25	lasso	7.511786	1.185245e-01	133	5.547033	7.215437	-77008.915	-76389.888
26	lasso	8.095989	8.699403e-02	200	6.944114	8.839126	-82654.109	-81716.715
27	lasso	8.406786	7.453267e-02	371	11.401896	14.472548	-103637.852	-101904.083
28	lasso	7.783674	8.967643e-02	155	5.499727	6.889915	-77492.838	-76768.737
29	lasso	7.918039	8.686267e-02	307	10.497262	13.044423	-89097.449	-87660.913
30	lasso	7.723920	1.173442e-01	156	6.478862	8.414825	-85077.282	-84349.050
31	lasso	7.721938	7.495275e-02	288	7.153558	8.993838	-61569.102	-60221.762
32	lasso	7.779939	2.287634e-01	116	11.089571	14.104839	-235586.895	-235047.163
33	lasso	7.601268	8.819841e-02	231	7.778746	9.426509	-57639.033	-56565.048
34	lasso	6.760260	1.102453e-01	180	7.494705	9.634630	-87539.443	-86700.8973
35	lasso	7.507537	1.387551e-01	160	9.095350	11.113775	-105141.150	-104394.1250
36	lasso	6.460025	9.165141e-02	190	6.635396	8.294781	-50653.354	-49767.7259
37	lasso	6.315596	8.454798e-02	219	6.368091	8.158957	-71980.973	-70963.0226
38	lasso	7.219430	9.373897e-02	283	9.709356	12.208572	-141244.987	-139923.2219
39	lasso	6.994804	1.015472e-01	233	8.362660	10.022822	-90138.402	-89050.1271
40	lasso	7.267626	1.103399e-01	213	9.150801	11.338479	-110979.325	-109980.6945
41	lasso	7.266552	9.881759e-02	215	7.728890	9.981163	-84468.977	-83464.6054
42	lasso	7.808802	1.057044e-01	292	11.176352	14.011061	-154544.069	-153183.772
43	lasso	7.391182	2.129165e-01	144	14.110311	17.088465	-247147.030	-246476.596
44	lasso	7.427056	9.467480e-02	183	7.036283	8.744447	-56974.222	-56123.682
45	lasso	5.964738	9.458849e-02	177	5.714158	7.210022	-72949.316	-72127.484
46	lasso	5.998578	8.664760e-02	242	8.283569	10.276108	-58121.653	-56991.160
47	lasso	6.035167	1.363115e-01	262	12.438872	15.630875	-169639.562	-168413.959
48	lasso	5.901440	8.118151e-02	202	5.710802	7.060636	-63875.484	-62931.380
49	lasso	5.661515	1.051555e-01	134	5.666596	7.271462	-67562.746	-66942.719
50	lasso	5.646002	8.620232e-02	148	4.905328	6.239515	-41666.332	-40981.227

Table 4: 'Results of the various measured variables from 50 simulations of Lasso Regression'

	model	time_elapsed	lambda	number_coef	MAE	MSE	AIC	BIC
1	ridge	1240.358124	76.37737649	19199	4.577924	5.732449	-12075.998	77787.050
2	ridge	1224.488790	55.29356064	19377	7.401527	9.266910	-2421.995	88201.174
3	ridge	1223.688273	129.54412516	19566	6.113517	7.697539	-101993.236	-10095.761
4	ridge	1233.025464	60.45175873	19324	4.840175	6.157070	1561.080	92321.873
5	ridge	1230.757103	60.24081295	19575	11.628498	14.341373	-28500.435	62725.663
6	ridge	1188.372025	9.573440e+03	19981	13.242310	16.097681	-29280.025	64933.335
7	ridge	1230.713304	6.817842e+01	19419	5.553667	7.178192	-11645.994	79344.054
8	ridge	1213.181114	1.067799e+04	19496	7.657577	9.858853	-38962.739	52436.736
9	ridge	1230.821732	1.023167e+02	19529	7.583565	9.617338	-56681.786	34553.673
10	ridge	1230.961438	6.250402e+01	19398	7.633242	9.691397	-7232.810	83779.539
11	ridge	1205.733263	8.986625e+03	19986	8.197722	10.270595	-25814.302	66913.617
12	ridge	1190.705234	9.284283e+03	19977	9.269075	11.300743	-10676.866	83052.198
13	ridge	1226.701488	6.104049e+01	19320	4.815305	6.028298	-3971.602	86090.770
14	ridge	1202.866297	8.264801e+03	19994	5.436283	6.888197	-16734.048	76444.831
15	ridge	1242.389871	5.608611e+01	19380	5.625884	7.126306	5235.510	96520.737
16	ridge	1196.929426	8.152836e+03	19987	8.643842	10.823035	-6964.862	86586.700
17	ridge	1196.041067	8.133461e+03	19978	7.570053	9.505521	-8509.072	85641.654
18	ridge	1197.645010	1.270006e+04	19998	8.695144	11.530727	-95253.947	-2210.645
19	ridge	1245.720435	5.489972e+01	19315	6.406334	8.076552	5059.931	95417.431
20	ridge	1202.760417	6.965613e+03	19982	6.345942	7.802051	6754.812	100308.024
21	ridge	1246.298311	7.485994e+01	19601	11.349403	14.663876	-47791.759	44270.113
22	ridge	1243.990734	7.274719e+01	19514	12.216426	14.821842	-51326.893	40229.737
23	ridge	1225.638833	7.834906e+03	19515	10.541363	13.457978	3180.047	94497.745
24	ridge	1245.439611	6.806947e+01	19487	8.553275	10.601359	-14914.306	76612.104
25	ridge	1206.740042	8.558199e+03	19994	5.873368	7.679135	-21373.882	72385.277
26	ridge	1256.316940	7.442268e+01	19419	6.826841	8.710366	-39160.605	52308.332
27	ridge	1260.049005	5.825512e+01	19429	10.819409	13.596109	-25971.755	65065.152
28	ridge	1258.930763	9.087033e+01	19477	5.583227	7.024172	-35071.895	56503.485
29	ridge	1260.737958	7.108993e+01	19486	10.009855	12.349825	-25304.706	66168.836
30	ridge	1242.289471	8.702421e+03	19502	6.601391	8.440998	-29515.450	62105.475
31	ridge	1255.622380	6.228841e+01	19514	6.681345	8.383995	-6180.689	85424.301
32	ridge	1216.207545	1.551970e+04	19997	12.069072	15.373066	-143367.143	-49519.443
33	ridge	1255.989272	6.293716e+01	19310	7.465640	8.991735	-9129.466	81033.872
34	ridge	1190.538415	8.773217e+03	19984	7.734656	10.047050	-19960.200	73652.3966
35	ridge	1189.933742	9.332673e+03	19992	8.962508	11.051981	-38432.600	55490.4767
36	ridge	1191.831337	6.529411e+03	19978	6.677882	8.432709	7324.726	100934.1700
37	ridge	1229.682136	8.310808e+01	19194	6.155322	7.933329	-27596.128	62025.5489
38	ridge	1226.861384	9.585350e+01	19610	9.266621	11.779783	-77230.492	14679.0206
39	ridge	1232.628528	7.986274e+01	19473	8.316984	10.022847	-41068.081	50271.9999
40	ridge	1194.522477	9.103082e+03	19984	9.561201	12.165108	-12200.219	81930.1643
41	ridge	1218.587630	8.087273e+03	19450	8.091087	10.293465	-5341.571	85938.8812
42	ridge	1224.270552	1.085531e+02	19572	10.661880	13.398492	-83416.678	8069.163
43	ridge	1178.444533	1.429559e+04	19999	14.338818	17.518364	-150781.316	-57023.598
44	ridge	1219.159161	7.087789e+01	19472	6.902654	8.571909	-12975.251	78018.534
45	ridge	688.022463	7.533408e+01	19533	5.578888	7.065082	-30214.101	60990.536
46	ridge	682.253289	6.848852e+01	19356	8.018833	10.005409	-11634.591	79156.661
47	ridge	641.275389	9.518068e+03	19977	11.547066	14.624845	-39912.279	53891.000
48	ridge	693.360931	8.179318e+01	19451	5.686671	6.934609	-21604.285	69753.031
49	ridge	642.705960	8.274295e+03	19989	6.284328	8.175173	-13871.492	79309.660
50	ridge	680.248218	6.628734e+01	19235	4.852651	6.137803	-1853.465	87788.038

Table 5: 'Results of the various measured variables from 50 simulations of Ridge Regression'

	model	time_elapsed	lambda	number_coef	MAE	MSE	AIC	BIC
1	blasso	348.485348	152.56434252	544	4.923115	6.255333	3515.212	6062.280
2	blasso	347.712912	171.74323280	547	7.889699	9.894517	3906.781	6465.834
3	blasso	365.518734	142.28270733	559	6.487014	8.009385	3776.927	6403.259
4	blasso	355.833789	234.41389670	573	5.218930	6.680190	3840.934	6533.041
5	blasso	324.494452	265.66780560	538	11.852203	14.715149	4198.913	6706.990
6	blasso	377.185492	2.523930e+02	566	13.572947	17.078889	4567.589	7237.182
7	blasso	350.724324	2.730728e+02	532	5.976567	7.704859	3741.509	6235.051
8	blasso	350.959795	2.508875e+02	545	8.028331	10.358256	4088.741	6644.573
9	blasso	339.967708	1.920565e+02	554	7.908912	10.138988	3900.278	6489.285
10	blasso	365.280201	2.396850e+02	564	8.119911	10.417195	4062.223	6709.258
11	blasso	325.760121	2.180464e+02	525	9.028594	11.173710	3773.619	6210.240
12	blasso	374.711706	2.108939e+02	550	8.950653	11.561759	4113.530	6694.858
13	blasso	355.194591	1.805751e+02	537	5.286062	6.844636	3580.169	6084.271
14	blasso	346.095196	2.772913e+02	546	5.902274	7.368489	3634.541	6179.914
15	blasso	405.575156	1.771188e+02	560	6.051638	7.623090	3779.248	6417.823
16	blasso	363.458786	1.638838e+02	557	8.690569	10.712210	3931.138	6539.073
17	blasso	391.606173	1.831952e+02	568	6.719150	8.799542	4029.031	6706.680
18	blasso	331.174500	2.262611e+02	549	9.272794	11.671211	4046.325	6601.455
19	blasso	350.798576	1.642327e+02	565	6.475497	8.188195	3802.593	6446.566
20	blasso	370.521561	1.960164e+02	564	6.125628	7.611865	3686.381	6327.796
21	blasso	377.737010	2.747676e+02	567	11.791684	15.456637	4367.888	7031.806
22	blasso	367.418967	2.357347e+02	553	12.664207	15.754405	4389.058	6984.466
23	blasso	352.783180	1.676919e+02	530	11.442544	14.388648	4168.700	6649.553
24	blasso	389.069071	2.322625e+02	568	9.029021	11.143959	4054.544	6723.160
25	blasso	375.345651	2.095706e+02	544	6.127861	7.983743	3790.193	6342.011
26	blasso	414.479290	2.339559e+02	569	7.406876	9.680804	3975.591	6656.571
27	blasso	367.477320	2.285084e+02	567	11.378481	14.237897	4287.787	6945.378
28	blasso	382.237248	1.787941e+02	533	5.968529	7.677030	3760.030	6266.830
29	blasso	390.020246	3.323593e+02	554	10.052005	12.525838	4207.479	6808.951
30	blasso	375.999975	2.136288e+02	556	6.928199	8.952149	3883.933	6496.855
31	blasso	377.196017	1.972470e+02	541	7.406497	9.248738	4002.496	6542.923
32	blasso	368.677477	1.946298e+02	547	12.011562	15.362620	4318.673	6886.599
33	blasso	343.724983	2.068780e+02	553	7.799459	9.537339	3899.376	6482.310
34	blasso	354.515120	1.796802e+02	539	7.845005	10.082675	3913.398	6439.0772
35	blasso	378.450752	2.722242e+02	555	9.400006	11.443980	4150.849	6759.0717
36	blasso	355.552502	2.416538e+02	554	6.731763	8.429011	3869.610	6466.2683
37	blasso	342.347272	1.895159e+02	542	6.733372	8.545690	3768.823	6300.3786
38	blasso	362.227241	1.588890e+02	553	10.308731	13.187981	4146.780	6739.4404
39	blasso	371.491334	1.016931e+02	552	9.104701	10.789705	3972.072	6562.1018
40	blasso	381.225772	1.684793e+02	550	9.302199	11.927476	4227.099	6818.5560
41	blasso	365.678731	2.562561e+02	578	7.996557	10.443296	4102.239	6815.6955
42	blasso	343.735734	2.297004e+02	550	11.430955	14.429847	4232.592	6804.294
43	blasso	358.150734	2.956631e+02	554	13.993811	17.397684	4408.136	7006.174
44	blasso	355.375004	2.319366e+02	539	7.171086	9.134572	3784.335	6303.924
45	blasso	315.294102	2.244371e+02	563	6.087522	7.554784	3737.092	6366.734
46	blasso	339.952037	2.653003e+02	563	8.182580	10.145017	3984.310	6625.953
47	blasso	330.182354	2.280173e+02	561	12.749643	16.265842	4463.346	7098.383
48	blasso	334.851322	2.808561e+02	570	5.934952	7.328664	3783.675	6461.687
49	blasso	309.593950	2.060911e+02	534	6.006868	7.724111	3642.043	6132.155
50	blasso	309.610264	1.448171e+02	548	5.379426	6.718698	3621.192	6175.888

Table 6: 'Results of the various measured variables from 50 simulations of Bayesian Lasso Regression'