Introduction to Statistical Machine Learning for Bioinformaticians and Medical Informaticians

SoSe 2022

Tutor: Dana Petracek

1	2	$\sum$

Marina Dittschar & Clarissa Auckenthaler

Assignment 6

(Submitted 20.06.2022)

Task 1

Derive the variance formula:

$$Var(\frac{1}{k}\sum_{i=1}^{k} X_i) = \rho\sigma^2 + \frac{1-\rho}{k}\sigma^2$$

where $X_i, i = 1, ..., k$, are identically distributed random variables with positive pairwise correlation ρ and $Var(X_i) = \sigma^2$ for $i = 1, ...k$.

The general formula for the variance is [1]:

$$Cov(X, X) = \frac{1}{k-1}\sum_{i=1}^{k}(X_i - \bar{X})(X_i - \bar{X}) \tag{1}$$

Therefore, when we introduce a multiplication factor of $\frac{1}{k}$ to the data, we get the following:

$$Cov(X, X) = \frac{1}{k-1}\sum_{i=1}^{k}\frac{1}{k}(X_i - \bar{X})(X_i - \bar{X})\frac{1}{k} \tag{2}$$

We will denote the vector resulting from every ith element $\frac{1}{k}$ for all i as κ. In the general case of an original covariance matrix Σ, this means that[2]:

$$Cov(\kappa X, \kappa X) = \kappa^T \Sigma \kappa \tag{3}$$

Now we know that these variables are identically distributed, but still correlated with positive pairwise correlation ρ for all pairs, i.e. i=1,...,k. They also all have the same variance of $Var(X_i) = \sigma^2$. We know that the covariance of two variables is the correlation ρ multiplied by the square root of the variance [4]. We can therefore write the covariance of any given pair as:

$$Cov(X_i, X_j) = \rho \cdot \sqrt{Var(X_i) \cdot Var(X_j)} = \rho \cdot \sqrt{\sigma^2 \cdot \sigma^2} = \rho \cdot \sigma^2 \tag{4}$$

With the given variances and covariances we can construct the full covariance matrix as [2]:

$$\Sigma = \begin{bmatrix} \sigma^2 & \sigma^2 & \dots & \rho \cdot \sigma^2 \\ \rho \cdot \sigma^2 & \sigma^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \cdot \sigma^2 \\ \rho \cdot \sigma^2 & \dots & \rho \cdot \sigma^2 & \sigma^2 \end{bmatrix} \tag{5}$$

We can also see this as a full matrix of ρ, except for the diagonal. The whole matrix is further multiplied with σ^2. We can write this as [3]:

$$\Sigma = \sigma^2(\rho 1_k 1_k^T + (1-\rho)I_B) \tag{6}$$

Where 1_B is a vector of length k where all entries are equal to 1 and I_B is the identity matrix of size k.

We combine this knowledge with our prior derivation of the covariance matrix if X is multiplied by κ [2]:

$$\kappa^T \Sigma \kappa = \kappa^T \sigma^2(\rho 1_k 1_k^T + (1-\rho)I_B)\kappa \tag{7}$$

$$= \sigma^2 \rho \kappa^T 1_k 1_k^T \kappa + \kappa^T I_B \kappa (1-\rho)\sigma^2 \tag{8}$$

$$= \sigma^2 \rho \cdot 1 + \frac{1}{k}(1-\rho)\sigma^2 \tag{9}$$

The red part of equation (8) can be simplified as shown because:

$$\kappa^T 1_k = 1_k^T \kappa = k \cdot \frac{1}{k} \tag{10}$$

The blue part of equation (8) holds because:

$$\kappa^T I_B \kappa = \kappa^T \cdot \kappa = \frac{1}{k^2} + ... + \frac{1}{k^2} = k \cdot \frac{1}{k^2} = \frac{1}{k} \tag{11}$$

Therefore we arrive at the desired formula:

$$Var(\frac{1}{k}\sum_{i=1}^{k} X_i) = \rho\sigma^2 + \frac{1-\rho}{k}\sigma^2 \tag{12}$$

q.e.d.

Task 2

a) **Read and normalize the data: use *read.table()* to load the data; column 9 is the output lpsa for the regression and column 10 determines whether this data entry belongs to the training set. Column 1 is just an index and should not be used for prediction. Normalize each input feature to a mean of 0 and a variance of 1. Split up the data set into training and test set respectively. For this subtask, you may include the code (with**

proper formatting) into the pdf report. Useful functions: mean(), sd(), and the MASS library.

We do the normalization on all numerical features: *lcavol, lweight, age, lbph, svi, lcp, gleason, pgg45*, these are in the dataframe the columns 1 to 8. We used the following code snippet for the normalizatin of the data:

```
        # normalization over column 1 to 8
        for(i in 1:8 ){
            data[,i] <- (data[,i] - mean(data[,i]))/ sd(data[,i])
        }

```

b) **Use LOOCV, 5- and 10-fold cross-validation on the training data set to estimate the test error of using linear regression to predict lpsa from all other features. Use the full training data set to train a linear regression model and compute the test error. Compare your estimates obtained from cross-validation to the error obtained from the test set and argue about your findings. Which of the methods is (theoretically) the fastest?**

The table 1 displays the computed test errors for the different methods. As we can see the linear regression model has the smallest test error: ~ 0.521 compared to the cross-validation values. This implies that the more data a model is trained on, the smaller the test error (the amount of training data gets smaller and the test error larger in the order: Linear regression -> LOOCV -> 10-fold cv -> 5-fold cv). If we just compare them we see that the LOOCV test error is the smallest of the used cross-validation methods with an test error of ~ 0.584.

K-Fold Cross Validation runs K times faster than Leave One Out cross-validation (LOOCV), because K-fold cross-validation repeats the train/test split K-times. So from this perspective the 10-fold cross-validation is the fastest method.

Model	**Test Error**
LOOCV	0.5839552
5-fold cross-validation:	0.6334426
10-fold cross-validation:	0.6087333
Linear regression	0.521274

Table 1: Test error for the LOOCV, 5-fold cross-validation, 10-fold cross-validation, linear regression model

c) **Use the training set to fit ridge regression models and generate a plot showing the values of the coefficients in relation to the parameter λ (cf. Figure 6.4, p. 238, ISLR). What can you observe?**

In Figure 1, each curve corresponds to the ridge regression coefficient estimate for one of the eight variables, plotted as a function of λ.On the left side of this plot λ is 0 and the corresponding ridge coefficient estimates are the same as the usual least squares estimates. For larger λ the the ridge coefficient estimates are basically zero. We can observe in Figure 1 that the coefficients decrease with rising λ on average.The coefficient for 1 = lcavol(black) decreases the most. Some individual coefficients may temporarily increase 7= gleason (yellow), 3 = age (green), 6= lcp (pink).

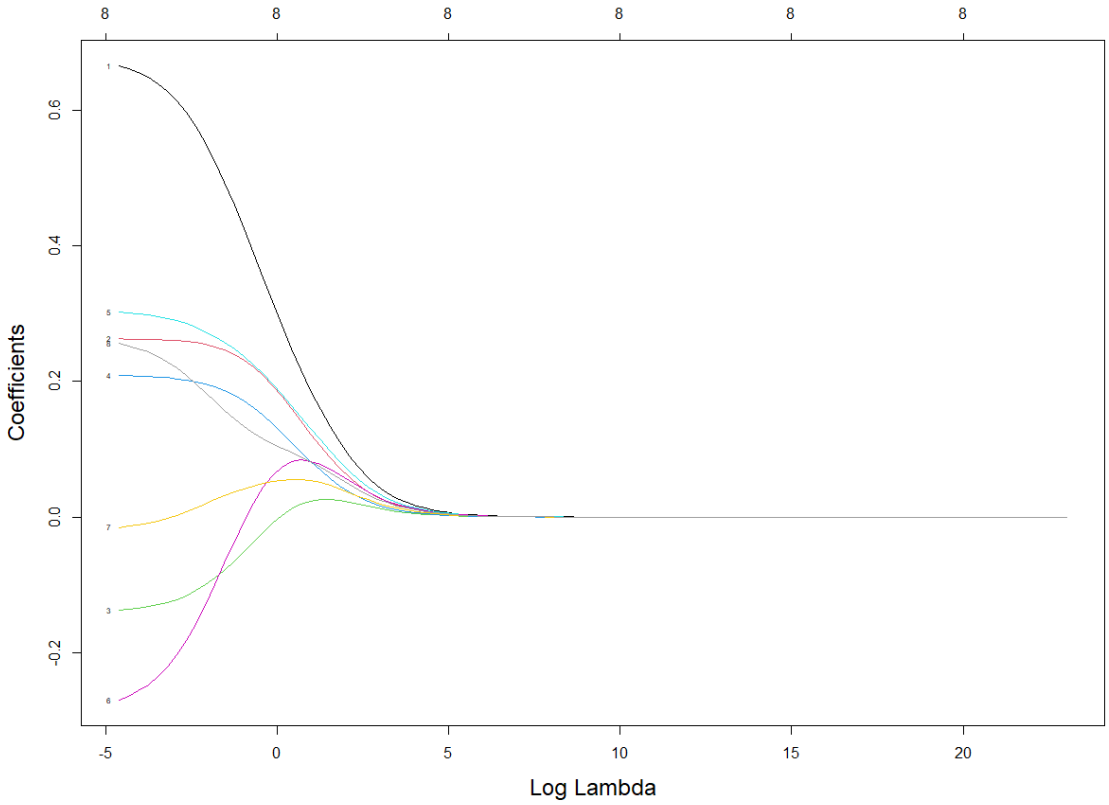

Figure 1: Fitted ridge regression model coefficients in relation to the parameter λ.

d) **Perform 10-fold cross-validation on the training set to determine the optimal value for λ for the ridge regression model. Report train and test error measured in MSE for this λ.**

To perform this we used the function `cv.glmnet()` that automatically performs k-fold cross validation using k = 10 folds. The optimal values are reported below:

Optimal lambda: $\lambda_{opt} = 0.08788804$
Train MSE = 0.4472905
Test MSE = 0.494454
Coefficients:

	s0
intercept	2.4670
lcavol	0.5807
lweight	0.2576
age	-0.1103
lbph	0.2002
svi	0.2813
lcp	-0.1632
gleason	0.0124
pgg45	0.1996

Table 2: Coefficient values for the ridge regression model, for the λ_{opt}

e) **Use the training set to fit lasso models and generate a plot showing the values of the coefficients in relation to the parameter λ (cf. Figure 6.6, p. 242, ISLR). What can you observe in comparison to the plot generated in (c)?**

The coefficient values for the lasso model are the same as for the ridge coefficients if $\lambda = 0$. In this case, the lasso simply gives the least squares fit, and when λ becomes sufficiently large, the lasso gives the null model in which all coefficient estimates equal zero. In between of these two extremes, the ridge regression and lasso models are quite different from each other:

Depending on the value of λ, the lasso can produce a model with any number of variables, whereas ridge regression will always include all of the variables in the model, although the magnitude of the coefficient estimates will depend on λ.

In Figure 2, each curve corresponds to the lasso coefficient estimate for one of the seven variables, plotted as a function of λ. The lasso coefficients decrease with rising λ on average. Some individual coefficients may temporarily increase (3= age (green), 6= lcp (pink). Compared to the plot in Figure 1 we can identify, that the coefficient 7= gleason is missing in Figure 2. The lasso coefficients converged faster to 0 compared to the plot of the ridge coefficients.

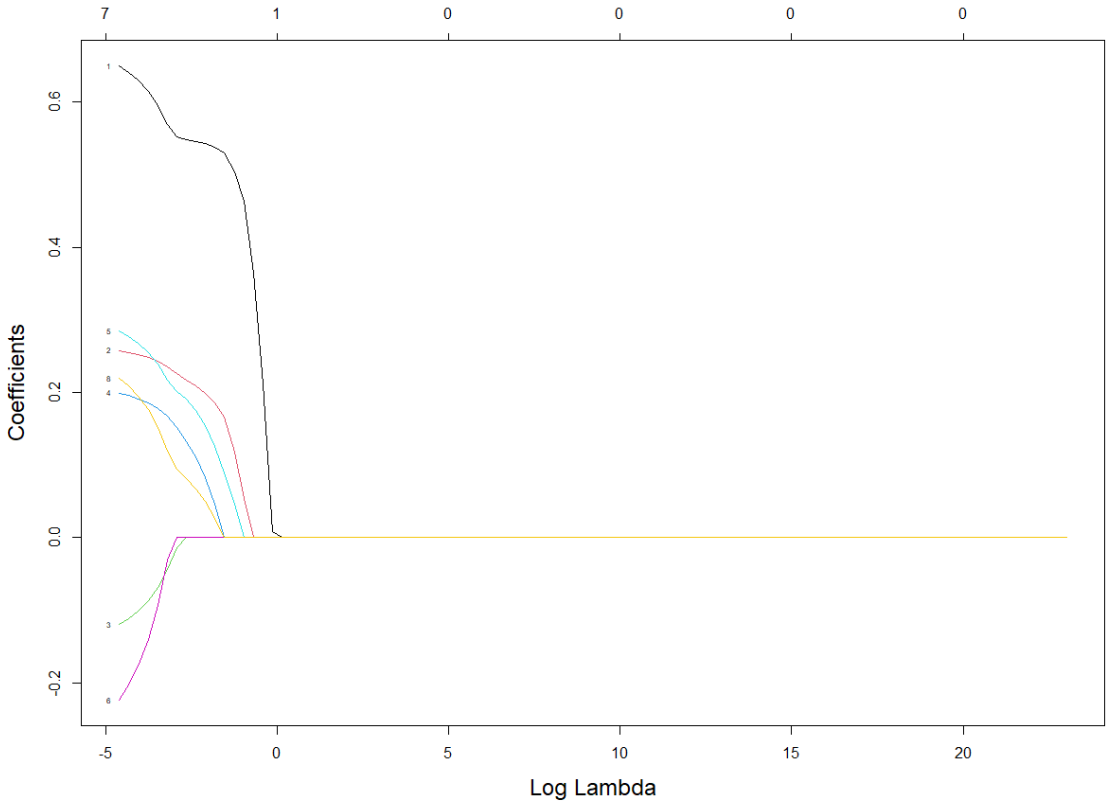

Figure 2: Fitted lasso model coefficients in relation to the parameter λ.

f) **Perform 10-fold cross-validation on the training set to determine the optimal value for λ in the lasso. Report train and test error measured in MSE for this λ. How many and which features are used? Compare this to the coefficients determined for ridge regression in (d).**

We again used the function `cv.glmnet()` that automatically performs k-fold cross validation using k = 10 folds.

Optimal lambda = $\lambda_{opt} = 0.007644054$
Train MSE = 0.441274
Test MSE = 0.4986605
Used features: 7 the values of the coefficients are displayed in table 3.
In the ridge regression we use all 8 features, because they can get close to zero but not zero, wheras with lasso some coefficients can "fall out" because they are getting zero, in this case our lasso model does not use the feature gleason.

Coefficients:

	s0
intercept	2.46703
lcavol	0.6500
lweight	0.2577
age	-0.1200
lbph	0.1991
svi	0.285
lcp	-0.2238
gleason	-
pgg45	0.2192

Table 3: Coefficient values for the lasso model, for the λ_{opt}

g) **Compare the models generated in (d) and (f) to the model generated in (a). Which model would you choose? What alternative model could have been used?**

Much like best subset selection, the lasso performs variable selection. As a result, models generated from the lasso are generally much easier to interpret than those produced by ridge regression.

If we compare the Test MSE of ridge regression and lasso (Table 4) we can see, that the test MSE for the ridge regression is slightly smaller than that of of the lasso. Both test MSE of ridge regression and lasso are smaller compared to the linear regression model. The estimated coefficients of the different models are all very close (Table 4). Based on the test MSE we would choose the lasso model.

Also we could have used the Elastic net regression model, it combines the properties of ridge and lasso regression.

Model	Df	%Def	Lambda	MSE_{Test}
ridge regression	8	68.87	0.08789	0.494454
lasso	8	69.43	0.007644	0.4986605
Linear regression				0.521274

Table 4: Comparison of the models (ridge regression, lasso, and linear regression) of the optimal λ

	linear regression	ridge regression	lasso
intercept	2.46530	2.4670	2.46703
lcavol	0.67953	0.5807	0.6500
lweight	0.26305	0.2576	0.2577
age	-0.14146	-0.1103	-0.1200
lbph	0.21015	0.2002	0.1991
svi	0.30520	0.2813	0.285
lcp	-0.28849	-0.1632	-0.2238
gleason	-0.02131	0.0124	-0.0167
pgg45	0.26696	0.1996	0.2192

Table 5: Comparison of the coefficient values of the models b), d), f)

References

[1] Nikolai Janakiev. Understanding the covariance matrix. `https://datascienceplus.com/understanding-the-covariance-matrix/`, 2018.

[2] Manuel Blum Prof. Dr. Martrin Riedmiller, Dr. Frank Hutter. Variance of decision trees. `https://ml.informatik.uni-freiburg.de/former/_media/teaching/ss14/sheet05_solution.pdf`, 2014.

[3] whuber (https://stats.stackexchange.com/users/919/whuber). What is the variance of the mean of correlated binomial variables? Cross Validated. URL:https://stats.stackexchange.com/q/83525 (version: 2014-01-27).

[4] Wikipedia. Kovarianz (stochastik). `https://de.wikipedia.org/wiki/Kovarianz_(Stochastik)#Zusammenhang_von_Kovarianz_und_Korrelation`, 2022.