Exercises – Lecture 4
Lasso regression

Question 1
Find the lasso regression solution for the data below for a general value of λ and for the straight line model $Y = \beta_0 + \beta_1 X + \varepsilon$ (only apply the lasso penalty to the slope parameter, not to the intercept). Show that when λ is chosen as 7, the lasso solution fit is $\hat{Y} = 40 + 1.75X$. Data: $\mathbf{X}^T = (X_1, X_2, \ldots, X_8)^T = (-2, -1, -1, -1, 0, 1, 2, 2)^T$, and $\mathbf{Y}^T = (Y_1, Y_2, \ldots, Y_8)^T = (35, 40, 36, 38, 40, 43, 45, 43)^T$.

Question 2
Consider fitting a multiple linear regression model by means of elastic net penalized least squares.

Question 2a)
Recall the data augmentation trick of Question 2 of the ridge regression exercises. Use the same trick to show that the elastic net least squares loss function can be reformulated to the form of the traditional lasso function. *Hint*: absorb the ridge part of the elastic net penalty into the sum of squares.

Question 2b)
The lasso can select maximally $\min\{n, p\} = rank(\mathbf{X})$ covariates. How many covariates can – in principle – the elastic net select?

Question 3
Investigate the effect of the variance of the covariates on variable selection by the lasso. Hereto consider the toy model: $Y_i = X_{1i} + X_{2i} + \varepsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, 1)$, $X_{1i} \sim \mathcal{N}(0, 1)$, and $X_{2i} = a\, X_{1i}$ with $a \in [0, 1]$. Draw a hundred samples for both X_{1i} and ε_i and construct both X_{2i} and Y_i for a grid of a's. Fit the model by lasso regression with $\lambda = 1$ for each choice of a. Plot e.g. in one figure *a)* the variance of X_{i1}, *b)* the variance of X_{2i}, and *c)* the indicator of the selection of X_{2i}. Which covariate is selected for which values of a?

Question 4
Augment the lasso penalty with the sum of the absolute differences all pairs of successive regression

coefficients:

$$\lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_F \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|.$$

This augmented lasso penalty is referred to as the *fused lasso penalty.*

Question 4a)
Consider the standard multiple linear regression model:

$$Y_i \quad = \quad \sum_{j=1}^{p} X_{ij}\, \beta_j + \varepsilon_i.$$

Estimation of the regression parameters takes place via minimization of penalized sum of squares, in which the fused lasso penalty is used with $\lambda_1 = 0$. Rewrite the corresponding loss function to the standard lasso problem by application of the following change-of-variables: $\gamma_1 = \beta_1$ and $\gamma_j = \beta_j - \beta_{j-1}$.

Question 4b)
Investigate on simulated data the effect of the second summand of the fused lasso penalty on the parameter estimates. In this, temporarily set $\lambda_1 = 0$.

Question 4c)
Let λ_1 equal zero still. Compare the regression estimates of Question 4b to the ridge estimates with a first-order autoregressive prior. What is qualitatively the difference in the behavior of the two estimates? *Hint:* plot the full solution path for the penalized estimates of both estimation procedures.

Question 4d)
How do the estimates of Question 4b change if we allow $\lambda_1 > 0$?

Question 5
A researcher has measured gene expression measurements for 1000 genes in 40 subjects, half of them cases and the other half controls.

Question 5a)
Describe and explain what would happen if the researcher would fit an ordinary logistic regression to these data, using case/control status as the response variable.

Question 5b)
Instead, the researcher chooses to fit a lasso regression, choosing the tuning parameter lambda by cross-validation. Out of 1000 genes, 37 get a non-zero regression coefficient in the lasso fit. In the ensuing publication, the researcher writes that the 963 genes with zero regression coefficients were found to be "irrelevant". What is your opinion about this statement?

Question 6
Download the `breastCancerNKI` package from BioConductor:

```
> source("http://www.bioconductor.org/biocLite.R")
> biocLite("breastCancerNKI")
```
Activate the library and load leukemia data from the package:
```
> library(breastCancerNKI)
> data(nki)
```
The eset-object `nki` is now available. It contains the expression profiles of 337 breast cancer patients. Each profile comprises expression levels of 24481 genes. Extract the expression data from the object, remove all genes with missing values, center the gene expression gene-wise around zero, and limit the data set to the first thousand genes. The reduction of the gene dimensionality is only for computational speed.
```
X <- exprs(nki)
X <- X[-which(rowSums(is.na(X)) > 0),]
X <- apply(X[1:1000,], 1, function(X) X - mean(X) ) .
```
Furthermore, extract the estrogen receptor status (short: ER status), an important prognostic indicator for breast cancer.
```
Y <- pData(nki)[,8]
```

Question 6a
Relate the ER status and the gene expression levels by a logistic regression model, which is fitted by means of ridge penalized maximum likelihood. First, find the optimal value of the penalty parameter of λ by means of cross-validation. This is implemented in `optL2`-function of the `penalized`-package available from `CRAN`.

Question 6b
Evaluate whether the cross-validated likelihood indeed attains a maximum at the optimal value of λ. This can be done with the `profL2`-function of the `penalized`-package available from `CRAN`.

Question 6c
Investigate the sensitivity of the penalty parameter selection with respect to the choice of the cross-validation fold.

Question 6d
Does the optimal lambda produce a reasonable fit? And how does it compare to the 'ridge fit'?

Answer to question 1
The design matrix is orthogonal. Hence, the lasso estimate of β_1 is unaffected by that of the intercept, and vice versa. Now first estimate the intercept by OLS and define: $\tilde{Y}_i = Y_i - \hat{\beta}_0$. This gives a modifierd lasso loss function:

$$(\tilde{Y}_i - \beta_1 X_i)^2 + \lambda_1 |\beta_1|.$$

Note that this is equivalent to optimization of:

$$c^2 \left\{ [\tilde{Y}_i/c - \beta_1 (X_i/c)]^2 + \frac{\lambda_1}{c^2} |\beta_1| \right\},$$

for $c > 0$. Or, equivalently:

$$[\tilde{Y}_i/c - \beta_1 (X_i/c)]^2 + \frac{\lambda_1}{c^2} |\beta_1|.$$

If we set $c = \|\mathbf{X}\|_2 = 4$, and write $\tilde{X}_i = X_i/c$, the design matrix is now orthonormal. We thus have an explicit solution for this simple lasso regression problem:

$$\hat{\beta}_1(\lambda) \quad = \quad \text{sgn}(\hat{\beta}^{OLS})(|\hat{\beta}^{OLS}| - \lambda)_+.$$

And thus $Y_i = 40 + (35 - \lambda) X_i/16$ for $\lambda < 35$, otherwise $Y_i = 40$. Substutition of λ yields the second part of the question.

Answer to question 2
Answer 2a
Write $\tilde{\mathbf{Y}} = (\mathbf{Y}^{\mathrm{T}}, \mathbf{0}_{1\times p})^{\mathrm{T}}$ and $\tilde{\mathbf{X}} = (\mathbf{X}^{\mathrm{T}}, \sqrt{\lambda_2}\,\mathbf{I}_{p\times p})^{\mathrm{T}}$. Then:

$$\|\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|_2^2 \quad = \quad \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda_2 \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{\beta}.$$

Hence, the elastic net loss function can be rewritten to:

$$\|\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1.$$

This is the traditional lasso problem.

Answer to question 2b
The rank of $\tilde{\mathbf{X}}$ equals p. Hence, in principle the elastic net can select all covariates.

Answer to question 3
Possible R-code one may use:

```
> # load library
> library(penalized)

> # draw data
> error <- rnorm(100)
> X1 <- rnorm(100)
> a <- c(0:200)/100
```

```
> # analysis
> coefs <- numeric()
> for (k in 1:length(a)){
+ X2 <- a[k] * X1
+ Y <- X1 + X2 + error
+ coefs <- rbind(coefs, coef(penalized(Y ~ X1 + X2, lambda1=1, unpenalized= 0), "all"))
+ }

> # calculate the variance of the covariates
> varX1 <- rep(var(X1), length(a))
> varX2 <- a * a * var(X1)

> # plot results
> plot(varX2, type="l", col="red", lwd=2, lty=1)
> lines(varX1, col="green", lwd=2, lty=2)
> lines((coefs[,2] != 0), lwd=2, col="black")
```

It should be clear from the plot that the second covariate is selected when its variance exceeds that of the first covariate (and vice versa). Hence, while both covariates have the same predictive power, the lasso selects one, the one with the largest variance.

Answer to question 4

Answer 4a

The loss function of the original problem is:

$$\sum_{i=1}^{n}(Y_i - X_{i,1}\beta_1 - X_{i_2}\beta_2 - \ldots - X_{i,p}\beta_p)^2 + \lambda_F \sum_{j=2}^{p} |\beta_j - \beta_{j-1}|.$$

The change-of-variables transforms this to:

$$\sum_{i=1}^{n}(Y_i - X_{i,1}\beta_1 - X_{i_2}\beta_2 - \ldots - X_{i,p}\beta_p)^2 + \lambda_F \sum_{j=2}^{p} |\gamma_j|.$$

Rests the rewrite the first summand. Hereto reformulate the change-of-variables to $\beta_j = \sum_{k=1}^{j} \gamma_k$. Substitute this:

$$\sum_{i=1}^{n}[Y_i - X_{i,1}\gamma_1 - X_{i_2}(\gamma_1 + \gamma_2) - \ldots - X_{i,p}(\gamma_1 + \ldots + \gamma_p)]^2 + \lambda_F \sum_{j=2}^{p} |\gamma_j|.$$

Reshuffling of terms and writing $U_{ij} = \sum_{k=j}^{p} X_{ik}$ gives:

$$\sum_{i=1}^{n}(Y_i - U_{i,1}\gamma_1 - U_{i_2}\gamma_2 - \ldots - U_{i,p}\gamma_p)^2 + \lambda_F \sum_{j=2}^{p} |\gamma_j|,$$

the traditional lasso problem.

Answer 4b
The following R-code code have used:

```
> # load library
> library(penalized)

> # set number of covariates
> p <- 3

> # create random data
> X <- matrix(rnorm(300), ncol=3)
> betas <- matrix(1:3, ncol=1)
> Y <- X %*% betas + matrix(rnorm(100, sd=0.25), ncol=1)

> # obtain the fussed lasso regression estimates
> lambdaF <- c(1:1500)/10
> betasHat <- numeric()
> for (k in 1:length(lambdaF)){
+ betasHat <- rbind(betasHat, coef(penalized(Y ~ X, unpenalized=~0,
+ fusedl=TRUE, lambda1=0, lambda2=lambdaF[k]), "all"))
+ }

> # plot ridge estimates verses lambdaF
> plot(betasHat[,1] ~ lambdaF, type="l", lwd=2, lty=1, col="red",
+ ylab="beta", xlab="rho", ylim=c(min(betasHat), max(betasHat)))
> lines(betasHat[,2] ~ lambdaF, type="l", lwd=2, lty=2, col="green")
> lines(betasHat[,3] ~ lambdaF, type="l", lwd=2, lty=3, col="blue")
```

Answer 4c
The ridge estimates converge to one another as ρ tends to one, but only in the limit they are equal. For the fused lasso, estimates may be equal long before $\lambda_F = \infty$. In particular, two estimates may be equal before they are 'merged' with the estimate of another regression parameter.

Answer 4d
If $\lambda_1 = 0$, the fused lasso estimates are not shrunken to zero. Eventually, as $\lambda_1 \to \infty$ the estimates vanish.

Answer to question 5
Answer 5a
There will complete separation between cases and controls in covariate space. Consequently, the estimates of the regression coefficients will go to $\pm\infty$, the fitted probabilities will go to 0 or 1, and the model will not converge.

Answer to question 5b
The lasso method introduces bias. This bias favors models in which a few covariates with large variance are selected to predict the response. The reason that variables that are not selected can reflect their lack of predictive ability, but also the bias in the lasso method. Variables may be left out because they have relatively small variance, or because they happen to be collinear to a stronger predictor. Stating that such variables are irrelevant is not warranted.

Answer to question 6
Code is similar to that of Question 7 of the ridge regression exercises, with e.g. `optL2` replaced by `optL1`.