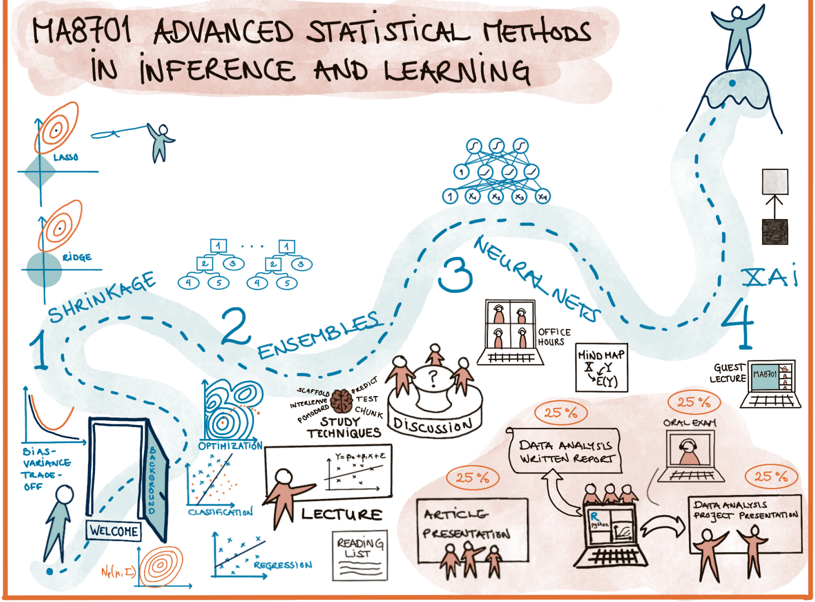
MA8701 Advanced methods in statistical inference and learning

L2: Shrinkage - the beginning

Mette Langaas IMF/NTNU

17 January, 2021

MA8701 ADVANCED STATISTICAL METHODS IN INFERENCE AND LEARNING



Shrinkage

Literature lecture 2 (L2)

- [ELS] The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics, 2009) by Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Ebook. Chapter 3.2 and 3.4.1-3.4.3.
- [HTW] Hastie, Tibshirani, Wainwrigh: "Statistical Learning with Sparsity: The Lasso and Generalizations". CRC press. Ebook. Chapter 1, 2.1-2.3,2.5.

and for the interested student

- Wessel N. van Wieringen: Lecture notes on ridge regression (We will refer to this note as WNvW below.)

Some figures are taken from An Introduction to Statistical Learning, with applications in R (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.

What is in a name?

This part of the course could have been called:

- "Regularized linear and generalized linear models"
- "Penalized maximum likelihood estimation"
- and also "Sparse models",

but it is called "Shrinkage".

Focus is on generalized linear models, but we will also consider shrinkage in the next parts of this course (then for "more complex" method).

Question: in linear models (linear regression, generalized linear regression) we mainly work with methods where parameter estimates are unbiased - but might have high variance and not give very good prediction performance overall. Can we use penalization (shrinkage) to produce parameter estimates with some bias but less variance, so that the prediction performance is improved?

We will look at different ways of penalization (which produces shrunken estimators) - mainly what is called ridge and lasso methods.

Ridge is not a sparse method, but lasso is. In sparse statistical models a *small number of covariates* play an important role.

HTW (page 2): *Bet on sparsity principle: Use a procedure that does well in sparse problems, since no procedure does well in dense problems.*

Shrinkage (penalization, regularization) methods are especially suitable in situations where we have multi-collinearity and/or more covariates than observations $N << p$. Two examples are

- in medicine with genetic data, where the number of patient samples is less than the number of genetic markers studied,
- in analysis of text (more to come in L3)

Linear models

(ELS 3.2, HTW Ch 2.1)

We will only consider linear models in L2, and move to generalized linear models in L3.

Set-up

Random response Y and p-dimensional (random) covariates X.
Training data: N (independent) observations: (y_i, x_i), where x_i is a column vector with p covariates (features).

Linear regression model

(ELS 3.2)
Additive noise model

$$Y = f(X) + \varepsilon$$

with $\mathrm{E}(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2$.
With squared loss, we remember that the optimal $f(X) = \mathrm{E}(Y \mid X)$.
Linear regression model - we assumes that

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

is linear in X, or that is a good approximation.
The unknown parameters are the regression coefficients $\beta_0, \ldots, \beta_p$ and the error variance σ_ε^2.
From TMA4267 we know that if (X, Y) is jointly multivariate normal, then the conditional distribution of $Y \mid X$ has mean that is linear in X and variance that is independent of X. Brush-up: See classnotes page 8.

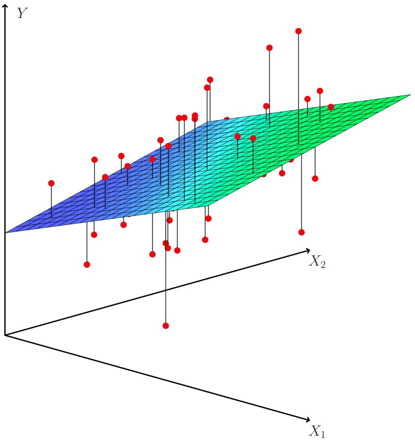


Figure 1: Figure from An Introduction to Statistical Learning, with applications in R (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.

Covariates

The covariates X can be both quantitative or qualitative, be made of basis expansions or interactions - and more. For qualitative covariates often a dummy variable coding is used. Brush-up: See TMA4315 GLM Module 2.

For now we don´t say so much more, but later we want the covariates to be standardized and the reponse to be centered.

The classical linear model and least squares estimation

First version

For the classical linear model we assume

$$Y_i = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \varepsilon_i$$

with $\mathrm{E}(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_) = \sigma_\varepsilon^2$, and independence of errors $\varepsilon_j, \varepsilon_i$.
Regression parameters $\beta = (\beta_0, \beta_1, \ldots, \beta_p) \in \Re^{(p+1)}$.
We will use the word *linear predictor* $\eta(x_i) = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j$, for the linear combination in the parameters β.
The least squares estimator for the parameters β is found by minimizing the squared-error loss:

$$\text{minimize}_\beta \{ \sum_{i=1}^{N} (y_i - \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j)^2 \}$$

Second version

This can also be written with vectors and matrices for the $i = 1, \ldots, N$ observations.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$$

where $\mathbf{Y}$ is a $N \times 1$ random column vector, $\mathbf{X}$ a $N \times (p+1)$ design matrix with row for observations and columns for covariates, and ε $N \times 1$ random column vector

The assumptions for the classical linear model is:

1. $\text{E}(\varepsilon) = \mathbf{0}$.
2. $\text{Cov}(\varepsilon) = \text{E}(\varepsilon\varepsilon^T) = \sigma^2\mathbf{I}$.
3. The design matrix has full rank, $\text{rank}(\mathbf{X}) = (p+1)$.

The classical *normal* linear regression model is obtained if additionally

4. $\varepsilon \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$ holds.

For random covariates these assumptions are to be understood conditionally on $\mathbf{X}$.

For derivation of the least squares estimator $\hat{\beta}$ see TMA4268 Module 3 and links therein.

The same results are found using likelihood theory, if we assume that $Y \sim N$. See TMA4315 GLM Module 2. Both methods are written out in these class notes from TMA4267/8.

The squared error loss to be minimized can be written

$$(\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)$$

Differensiation with respect to the unknown parameter vector, and equating to zero leads to the *normal equations*.

$$\mathbf{X}^T\mathbf{X}\beta = \mathbf{X}^T\mathbf{Y}$$

To give

$$\hat{\beta}_{\mathsf{L}S} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Properties of estimators

If we only assume a classical linear model, the mean and covariance of $\hat{\beta}$ is $\mathrm{E}(\hat{\beta}_{\mathrm{L}S}) = \beta$ and $\mathrm{Cov}(\hat{\beta}_{\mathrm{L}S}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.

For the classical normal linear model:

- Least squares and maximum likelihood estimator for β:

  $$\hat{\beta}_{\mathrm{L}S} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

  with $\hat{\beta}_{\mathrm{L}S} \sim N_p(\beta, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$.
- Restricted maximum likelihood estimator for σ^2:

  $$\hat{\sigma}^2 = \frac{1}{n-p}(\mathbf{Y} - \mathbf{X}\hat{\beta}_{\mathrm{L}S})^T(\mathbf{Y} - \mathbf{X}\hat{\beta}_{\mathrm{L}S}) = \frac{\mathrm{SSE}}{n-p}$$

  with $\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$.
- Statistic for inference about β_j, c_{jj} is diagonal element j of $(\mathbf{X}^T\mathbf{X})^{-1}$.

  $$T_j = \frac{\hat{\beta}_{\mathrm{L}S,j} - \beta_j}{\sqrt{c_{jj}}\hat{\sigma}} \sim t_{n-p-1}$$

The Gauss-Markov theorem

(ELS 3.2.2)
The Gauss-Markov theorem is the famous result stating: *the least squares estimators for the regression parameters β have the smallest variance among all linear unbiased estimators*.
For simplicity, we look at a linear combination of the parameters, $\theta = a^T\beta$, with estimator $\hat{\theta} = a^T\hat{\beta} = a^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. Observe that the estimator is linear in the response $\mathbf{Y}$.
Q: why is a linear combination of interest? What about a prediction of the response at covariate x_0? It would be $f(x_0) = x_0^T\beta$, a linear combination of the β elements.

If we assume that the linear model is correct, then $\hat{\theta}$ is an unbiased estimator of θ, because $\mathrm{E}(a^T\hat{\beta}) = a^T\mathrm{E}(\hat{\beta}) = a^T\beta = \theta$.

According to the Gauss-Markov theorem: if we have another estimator $\tilde{\theta} = c^T\mathbf{Y}$ that is unbiased for θ then it must have a larger variance than the LS-estimator:

$$\mathrm{Var}(\hat{\theta}) = \mathrm{Var}(a^T\hat{\beta}) \leq \mathrm{Var}(c^T\mathbf{Y}) = \mathrm{Var}(\tilde{\theta})$$

In Exercise ELS 3.3a we prove the Gauss-Markov theorem based on this set-up (least squares estimator of a linear combination $a^T\beta$).

Proof for the full parameter vector β (not only the scalar linear combination), requires a bit more work (it is ELS exercise 3.3b if you want to try).

Comparing variances of estimators

It is not hard to check that an estimator (for example $p \times 1$ column vector) is unbiased (in each element).

But, what does it mean to compare the variance (covariance matrix) of two estimators of dimension $p \times 1$?

In statistics a “common” strategy is to consider all possible linear combinations of the elements of the parameter vector, and check that the variance of estimator $\hat{\beta}$ is smaller (or equal to) the variance of another estimator $\tilde{\beta}$.

This is achieved by looking at the difference between the covariance matrices $\text{Cov}(\tilde{\beta}) - \text{Cov}(\hat{\beta})$. If the difference is a positive semi-definite matrix, then every linear combination of $\hat{\beta}$ will have a variance that is smaller or equal to the variance of the corresponding linear combination for $\tilde{\beta}$.

Why is this correct?

Assume we want to see if $\mathrm{Var}(c^T\tilde{\beta}) \geq \mathrm{Var}(c^T\hat{\beta})$ for any (nonzero) vector c.
We know that $\mathrm{Var}(c^T\hat{\beta}) = c^T\mathrm{Cov}(\hat{\beta})c$ and $\mathrm{Var}(c^T\tilde{\beta}) = c^T\mathrm{Cov}(\tilde{\beta})c$.
We then consider

$$\mathrm{Var}(c^T\tilde{\beta}) - \mathrm{Var}(c^T\hat{\beta}) = c^T(\mathrm{Cov}(\tilde{\beta}) - \mathrm{Cov}(\hat{\beta}))c$$

If Cov($\tilde{\beta}$) $-$ Cov($\hat{\beta}$) is positive semi-definite then the variance difference will be equal or greater than 0 - by the definition of a positive semi-definite matrix.

This is also referred to as: The variance of $\tilde{\beta}$ exceeds *in a positive definite ordering sense* that of $\hat{\beta}$, and written Var($\tilde{\beta}$) $\succeq$ Var($\hat{\beta}$). (Remark: here both Var and Cov is used as notation for the variance-covariance matrix.)

Mean squared error

We want to study the mean squared error for the (scalar) estimator $\tilde{\theta}$.

From the previous section we know that $\tilde{\theta}$ could for example be the prediction at at covariate x_0? It would be $\tilde{\theta} = f(x_0) = x_0^T \beta$, and then $\text{MSE}(\tilde{\theta})$ would be an interesting quantity.

$$\text{MSE}(\tilde{\theta}) = \text{E}[(\tilde{\theta} - \theta)^2] = \text{Var}(\tilde{\theta}) + [\text{E}(\tilde{\theta}) - \theta]^2$$

The last transition: add and subtract $\text{E}(\tilde{\theta})$.

The first term is the variance, and the second the squared bias.

(There is no irredusible error since we are not considering a new observation, but we may of cause do that and add the irreducible error.)

We know that for unbiased estimators (bias equal to 0), the MSE will be the smallest for the LS-estimator. This means that if we want to try to get a lower MSE we can´t do that with an unbiased estimator!

This is a bit unusual to many of us, since we from our first course in statistics have been told about the glory of unbiased estimators!

But, if we shrink some of the regression coefficients towards 0, or set them equal to 0, then we get a *biased estimate* for the regression parameters. Biased estimates are the core of this part of the course. We may want to pay the price of a biased estimate with the gain of decreased variance, so that the MSE for might get lower than for the LS-estimate.

Preparing for shrinkage

Standarization of covariates

For shrinkage methods it is common to *standardize* the covariates, where standardize means that

- the covariates are first centered, that is $\frac{1}{N}\sum_{i=1}^{N} x_{ij} = 0$ for all $j = 1, \ldots, p$,
- and then scaled to unit variance, that is $\frac{1}{N}\sum_{i=1}^{N} x_{ij}^2 = 1$.

This is done in practice by first subtracting the mean and then dividing by the standard deviation. The standarization is only needed if the covariates are of different units or scales, because for shrinkage we will (for some of the method) penalize the optimization with the same penalty for all covariates.

Centering covariates and response

The intercept term β_0 will not be the aim for shrinkage in shrinkage methods.

To make the presentation of the shrinkage methods easier to explain and write down, HTW use the common trick to center all covariates *and* the response.

By centering the covariates and the response we may imagine moving the centroide of the data to the origin, where we do not need an intercept to capture the best linear regression hyperplane.

When both covariates and responses are centred the LS estimate for the intercept β_0 will be $\hat{\beta}_0 = 0$.

If interpretation is to be done for uncentred data we may calculate the estimated β_0 for uncentered data from the estimated regression coefficients and the mean of the original covariates and respons.

When covariates and responses are centred HTW remove β_0 from the regression model for the shrinkage methods. We will also do that.

Group discussion:

1) Why is the LS estimate $\hat{\beta}_0 = 0$ for centred covariates and centred response in the multiple linear regression model?
2) Explain what is done in the analysis of the Gasoline data directly below.

Choose yourself if you want to focus 1 or 2.

Gasoline data

Consider the multiple linear regression model, with response vector $\mathbf{Y}$ of dimension $(N \times 1)$ and p covariates and intercept in $\mathbf{X}$ $(N \times p + 1)$.

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \qquad (1)$$

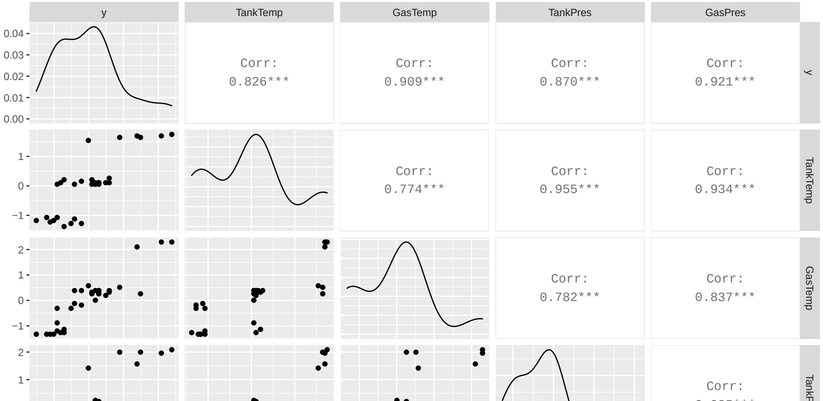
where $\varepsilon \sim \mathbf{N}(\mathbf{0}, \sigma^{\mathbf{2}}\mathbf{I})$.

When gasoline is pumped into the tank of a car, vapors are vented into the atmosphere. An experiment was conducted to determine whether Y, the amount of vapor, can be predicted using the

```
ds <- read.table("./sniffer.dat",header=TRUE)
x <- apply(ds[,-5],2,scale)
y <- ds[,5]-mean(ds[,5])
print(dim(x))
```

```
## [1] 32  4
```

```
dss=data.frame(y,x)
ggpairs(dss)
```



We have fitted a MLR with all four covariates. Explain what you see.

```
full <- lm(y~.,dss)
summary(full)
```

```
## 
## Call:
## lm(formula = y ~ ., data = dss)
## 
## Residuals:
##    Min     1Q Median     3Q    Max 
## -5.586 -1.221 -0.118  1.320  5.106 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)   
## (Intercept)  4.851e-15  4.826e-01   0.000  1.00000   
## TankTemp    -5.582e-01  1.768e+00  -0.316  0.75461   
## GasTemp      3.395e+00  1.065e+00   3.187  0.00362 **
## TankPres    -6.274e+00  4.140e+00  -1.515  0.14132   
## GasPres      1.249e+01  3.859e+00   3.237  0.00319 **
## ---
```

Perform best subset selection using Mallows C_p (equivalent to AIC) to choose the best model.

```r
bests <- regsubsets(x,y)
sumbests <- summary(bests)
print(sumbests)
```

```
## Subset selection object
## 4 Variables  (and intercept)
##           Forced in Forced out
## TankTemp      FALSE      FALSE
## GasTemp       FALSE      FALSE
## TankPres      FALSE      FALSE
## GasPres       FALSE      FALSE
## 1 subsets of each size up to 4
## Selection Algorithm: exhaustive
##           TankTemp GasTemp TankPres GasPres
## 1  ( 1 ) " "      " "     " "      "*"
## 2  ( 1 ) " "      "*"     " "      "*"
## 3  ( 1 ) " "      "*"     "*"      "*"
## 4  ( 1 ) "*"      "*"     "*"      "*"
```

Model after best subset selection.

```r
red <- lm(y~GasTemp+TankPres+GasPres,data=dss)
summary(red)
```

```
## 
## Call:
## lm(formula = y ~ GasTemp + TankPres + GasPres, data = ds
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -5.6198 -1.2934 -0.0496  1.4858  4.9131 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)   
## (Intercept)  5.131e-15  4.748e-01   0.000  1.00000   
## GasTemp      3.290e+00  9.951e-01   3.306  0.00260 **
## TankPres    -7.099e+00  3.159e+00  -2.247  0.03272 * 
## GasPres      1.287e+01  3.607e+00   3.568  0.00132 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.
```

Ridge regression

(ELS 3.4.1)

Ridge regression is also called “Tikhonov regularization”.

We consider the classical linear model set-up, as for the LS estimation, but now we look at shrinking the coefficients towards 0 to construct biased estimators - and then “hope” that this also has made the variances decrease.

We will not shrink the intercept β_0, because then the this will depend on the origin of the response.

The ridge solution is dependent on the scaling of the covariates, and usually we work with standardized covariates and also with centered response.

Minimization problem

Budget version

We want to constrain the size of the estimated regression parameters, so we give the sum of squared regression coefficients a budget t.
Minimize the squared error loss

$$\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

subject to $\sum_{j=1}^{p} \beta_j^2 \leq t$. The solution is called $\hat{\beta}_{\text{ridge}}$.

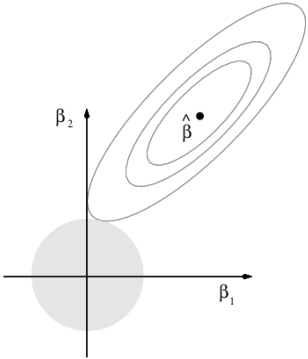


Figure 2: Figure from An Introduction to Statistical Learning, with applications in R (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.

Penalty version

$$\hat{\beta}_{\text{ridge}} = \text{argmin}_{\beta}[\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2]$$

where $\lambda \leq 0$ is a complexity (regularization, penalty) parameter controlling the amount of shrinkage.

- The larger λ the greater the amount of shrinkage
- The shrinkage is towards 0

This version of the problem is also called the Lagrangian form. The budget and penalty minimization problems are equivalent ways to write the ridge regression and there is a one-to-one correspondence between the budget t and the penalty λ.

Parameter estimation

As explained, centred covariates and responses are used - and the intercept term is removed from the model. Then NOW **X** does not include a column with 1s and has dimension $N \times p$.
Penalty criterion to minimize

$$(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta) + \lambda\beta^T\beta$$

This can be rewritten as

$$\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\beta + \beta^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\beta$$

Proceeding along the lines as done with the LS estimation, we get the (new) normal equations

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\beta = \mathbf{X}^T\mathbf{Y}$$

and the estimator:

$$\hat{\beta}_{\text{ridge}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}$$

Observe that the solution adds a positive constant λ to the diagonal of $\mathbf{X}^T\mathbf{X}$, so that even if $\mathbf{X}^T\mathbf{X}$ does not have full rank then the problem is non-singular and we can invert $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})$.

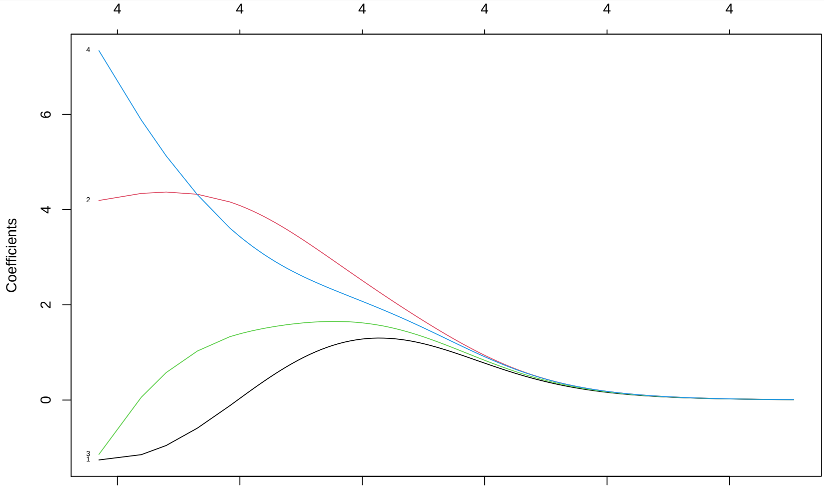
When ridge regression was introduced in statistics in the 1970s this (avoiding non-singuarlity) was the motivation.

When $N < p$ then the design matrix will have rank less than the number of covariates, and the LS estimate does not exist.

The case when two or more covariates are perfectly linearly dependent is called *super-collinearity* (accoring to WNvN).

Gasoline continued

```
start=glmnet(x=x,y=y,alpha=0)
autolambda=start$lambda # automatic choice of lambda had s
newlambda=c(autolambda,0.5,0.3,0.2,0.1)
fit.ridge=glmnet(x,y,alpha=0,lambda=newlambda)
plot(fit.ridge,xvar="lambda",label=TRUE)
```



Model selection

To choose the optimal penalty parameter λ cross-validation is the default method in use. ELS recommends to either

- choose the λ corresponding to the smallest CV error
- or first find the λ with the smallest CV-error, and then record the estimated standard error of the CV-error at this value, and then choose the largest λ such that the CV error is still within one standard error of the minimum. We choose the largest because we want the less flexible model.

The R package `glmnet` (by Hastie et al) has default $K = 10$ fold cross-validation with the function `cv.glmnet` where `alpha=0` gives the ridge penalty.

Gasoline continued

Explain what you see!

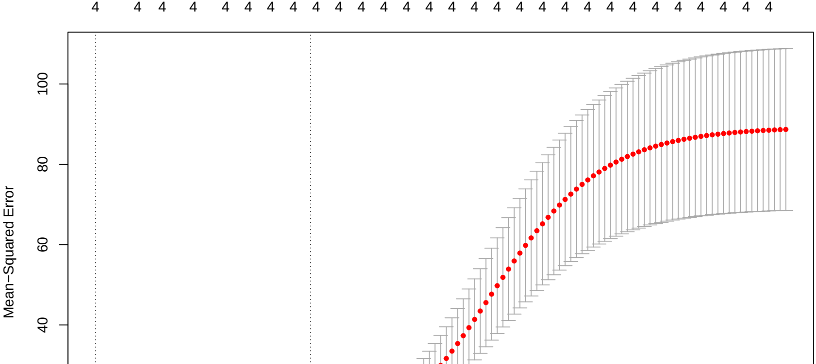
```
cv.ridge=cv.glmnet(x,y,alpha=0,lambda=newlambda)
print(paste("The lamda giving the smallest CV error",cv.ri
```

```
## [1] "The lamda giving the smallest CV error 0.1"
```

```
print(paste("The 1sd err method lambda",cv.ridge$lambda.1se
```

```
## [1] "The 1sd err method lambda 3.43009811945866"
```

```
plot(cv.ridge)
```



Properties of the ridge estimator

Mean

Derive the mean of the ridge estimator.
What happens if:

- $\lambda \to 0$
- $\lambda \to \infty$

Exam problem 12 (TMA4268, 2019) with solutions Alternatively: Wessel N. van Wieringen: Lecture notes on ridge regression, section 1.4

Covariance

Derive the covariance of the ridge estimator.

What happens if:

- $\lambda \to 0$
- $\lambda \to \infty$

(in our centered model without intercept)

Same resources as above.

Distribution

For the normal linear model

$$\hat{\beta}(\lambda)_{\text{ridge}} \sim N\{(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{X}\,\beta,$$
$$\sigma^2(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\}.$$

Is ridge "better than" LS?

- We may prove that the variance of the ridge estimator is smaller or equal the variance of the LS estimator. See exercise "Variance of ridge compared to LS", where we need to look at differences of covariance matrices and check for positive semi-definite matrix.
- In addition it is possible to prove that given a suitable choice for λ the ridge regression estimator may outperform the LS estimator in terms of the MSE. See WNvW Section 1.4.3 for the full derivation.
- The optimal choice of λ depends both the true regression parameters and the error variance. This means that the penalty parameter should be chosen in a *data-driven* fashion.

Insight based on SVD

Singular value decomposition (SVD)

Let $\mathbf{X}$ be a $N \times p$ matrix.
SVD is a decomposition of a matrix $\mathbf{X}$ into a product of three matrices

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T.$$

$\mathbf{D}$ is an $(N \times p)$-dimensional block matrix. Its upper left block is a $(\text{rank}(\mathbf{X}) \times \text{rank}(\mathbf{X}))$-dimensional digonal matrix with the singular values on the diagonal. The remaining blocks, zero if $p = N$. The singular values are equal $\sqrt{\mathrm{eigenvalues}(\mathbf{X}\mathbf{X}^T})= \sqrt{\mathrm{eigenvalues}(\mathbf{X}^T\mathbf{X}})$.
$\mathbf{U}$ is an $(n \times n)$-dimensional matrix with columns containing the left singular vectors (denoted $\mathbf{u}_i$), that is, the eigenvectors of $\mathbf{X}\mathbf{X}^T$
$\mathbf{V}$ is a $(p \times p)$-dimensional matrix with columns containing the right singular vectors (denoted $\mathbf{v}_i$), that is, the eigenvectors of $\mathbf{X}^T\mathbf{X}$.
The columns of $\mathbf{U}$ and $\mathbf{V}$ are orthogonal: $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_N = \mathbf{U}\mathbf{U}^T$ and $\mathbf{V}^T\mathbf{V} = \mathbf{I}_p = \mathbf{V}\mathbf{V}^T$.

Following the derivation of WNvW page 11-12:

- If $n > p$ and the rank of $\mathbf{X}$ is p, then the LS estimator $\hat{\beta}_{\text{LS}}$ can be written

$$\hat{\beta}_{\text{LS}} = \mathbf{V}(\mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{U}^T\mathbf{Y}$$

- The ridge estimator $\hat{\beta}_{\text{ridge}}$

$$\hat{\beta}_{\text{ridge}} = \mathbf{V}(\mathbf{D}^T\mathbf{D} + \lambda\mathbf{I})^{-1}\mathbf{D}^T\mathbf{U}^T\mathbf{Y}$$

- The principal component regression based on the first k principal components

$$\hat{\beta}_{\text{PCR}} = \mathbf{V}_k(\mathbf{I}_{kp}\mathbf{D}^T\mathbf{D}\mathbf{I}_{pk})^{-1}\mathbf{I}_{kp}\mathbf{D}^T\mathbf{U}^T\mathbf{Y}$$

here $\mathbf{V}_k$ contains the first k right singular vectors as columns, and $\mathbf{I}_{kp}$ is obtained by $\mathbf{I}_p$ by removing the last $p - k$ columns.

Connection to principal component analysis: The estimated covariance matrix for centred covariates is $\frac{1}{N}\mathbf{X}^T\mathbf{X}$. The eigenvalues of $\mathbf{X}^T\mathbf{X}$ are the squared singular values, d_j^2. The small singular values d_j correspond to directions in the column space of $\mathbf{X}$ with small variance, which will be the direction for the last principal components.

The ridge penalties shrinks the direction with the small singular values the most. Principal components thresholds coefficients in the direction with singular values of $\mathbf{X}$, while ridge regression shrinks the coefficients in these directions.

Alternatively, it is possible to consider the prediction

$$\hat{y}_{\mathrm{LS}} = \mathbf{X}\hat{\beta}_{\mathrm{L}S} = \cdots = \mathbf{U}\mathbf{U}^T\mathbf{y}$$

$$\hat{y}_{\mathrm{ridge}} = \mathbf{X}\hat{\beta}_{\mathrm{r}idge} = \cdots = \mathbf{U}\mathbf{D}^2(\mathbf{D}^2 + \lambda\mathbf{I}_p)^{-1}\mathbf{U}^T\mathbf{y} = \sum_{j=1}^{p} \mathbf{u}_j \frac{d_j^2}{d_j^2 + \lambda}\mathbf{u}_j^T\mathbf{y}$$

Group discussion: What can we conclude from this about what the λ does with each covariate direction?

The effective degrees of freedom

In ELS Ch 7.6 we defined the effective number of parameters (here now referred to as the *effective degrees of freedom*) for a linear smoother $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$ as

$$\mathrm{df}(\mathbf{S}) = \mathrm{trace}(\mathbf{S})$$

For ridge regression our linear smoother is

$$\mathbf{H}_\lambda = \mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T$$

$$\mathrm{df}(\lambda) = \mathrm{tr}(\mathbf{H}_\lambda) = \mathrm{tr}(\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T) = \cdots = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2+\lambda}$$

- $\lambda = 0$ gives $\mathrm{df}(\lambda) = p$
- $\lambda \to \infty$ gives $\mathrm{df}(\lambda) \to 0$

The $\mathrm{df}(\lambda)$ is sometimes plotted instead of λ on the horisontal axis when model complexity is chosen.

Finally

- When is ridge preferred to LS? When the LS estimates have high variance and many predictors are truly non-zero.
- Ridge is computationally fast.
- Ridge is not very easy to interpret, because all p predictor are included in the final model.

Lasso

(ELS 3.4.2)

Now we will do what looks at first sight as a small change - we will use a budget on the absolute value insted of squared value - moving from the L_2 to the L_1 norm. But, this will have a large impact on the parameter estimates - both shrinking - and performing model selection (by shrinking all the way down to 0).

Again, we will not shrink the intercept β_0, because then the this will depend on the origin of the response, and we will work with standardized covariates and centered response.

Minimization problem

Budget version

We want to constrain the size of the estimated regression parameters, so we give the sum of squared regression coefficients a budget t.
Minimize the squared error loss

$$\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

subject to $\sum_{j=1}^{p}|\beta_j| \leq t$. The solution is called $\hat{\beta}_{\text{lasso}}$.

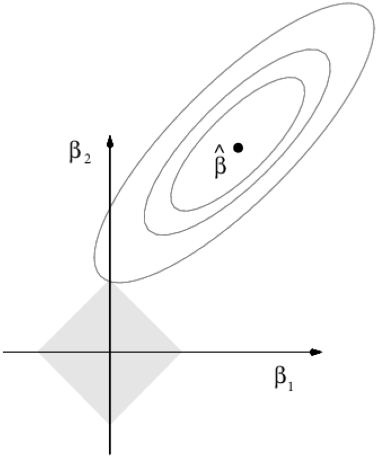


Figure 3: Figure from An Introduction to Statistical Learning, with applications in R (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.

Penalty version

$$\hat{\beta}_{\text{lasso}} = \text{argmin}_{\beta}[\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p}|\beta_j|]$$

again, $\lambda \leq 0$ is a complexity (regularization, penalty) parameter controlling the amount of shrinkage.

- The larger λ the greater the amount of shrinkage
- The shrinkage is towards 0

This version of the problem is also called the Lagrangian form. The budget and penalty minimization problems are equivalent ways to write the ridge regression and there is a one-to-one correspondence between the budget t and the penalty λ.

Small notational difference in the two textbooks

In HTW an extra $\frac{1}{2N}$ factor is added to the squared error for the ridge and the lasso, which is just for ease of interpretation of a future shrinkage parameter to be included (to make that shrinkage parameter comparable across different sample sizes in the use of cross-validation). The factor does not influence the solution of the minimization of the squared-error loss we consider now.

Parameter estimation

- As explained, centred covariates and responses are used - and the intercept term is removed from the model. Then $\mathbf{X}$ does not include a column with 1s and has dimension $N \times p$.
- The use of the absolute value in the penalty term makes the solution in general non-linear in y_i, and no closed form solution.
- If we make the budget t sufficiently small some of the coefficients will be exactly zero.
- If t is chosen larger than $t_0 = \sum_{j=1}^{p} |\hat{\beta}_{\text{LS},j}|$ the lasso estimates equal the LS estimates.
- The nature of the shrinkage is complex, and will be studied later.

In L3 we will look into estimation algorithms for the lasso.

Conditions for a solution to the penalty version

(HTW page 9)
The details are found in HTW Chapter 5 (not on our reading list), but the student familiar with convex analysis, dual problems and Karush-Kuhn-Tucker (KKT) conditions might find Chapter 5 of interest.
Convex analysis theory: necessary and sufficient conditions for a solution to the lasso penalty problem is

$$\frac{1}{N}\langle \mathbf{x}_j, \mathbf{y} - \mathbf{X}\beta \rangle + \lambda s_j = 0 \text{ for } j = 1, \ldots, p$$

where $\langle a, b \rangle = a^T b$ denotes the inner product. Each s_j is an unknow quantity, equal to

- sign(β_j) if $\beta_j \neq 0$
- some value in $[-1, 1]$ otherwise (socalled *subgradient* of the absolute value function).

We may solve this problem in $(\hat{\beta}, \hat{s})$, instead of the penalty version.

Orthogonal covarates

This case - explicit solution! New word: soft thresholding".

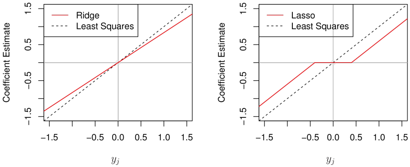
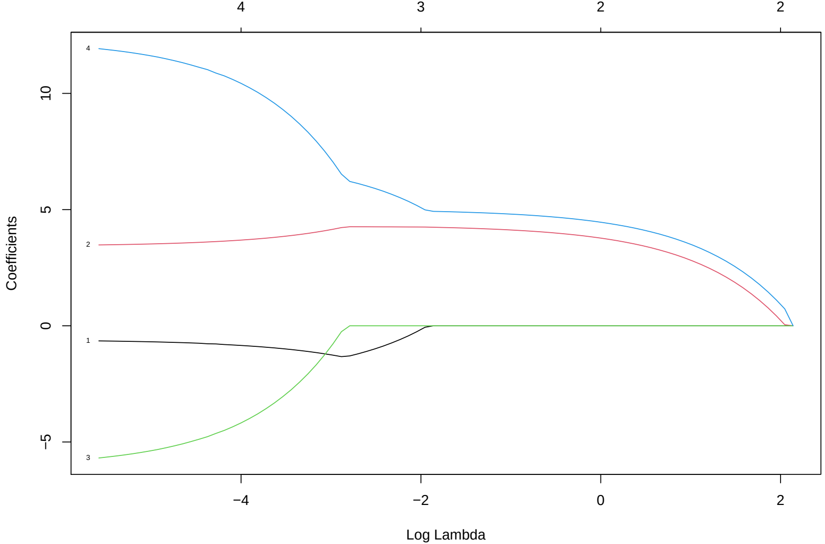
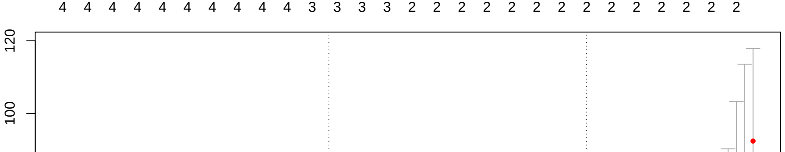


Figure 4: Figure from An Introduction to Statistical Learning, with applications in R (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani.

Gasoline continued





Degrees of freedom

(HTW 2.5)

In ELS Ch 7.6 we defined the effective number of parameters (here now referred to as the *effective degrees of freedom*) for a linear smoother, and used that for the ridge regression. However, the lasso is not a linear smoother (it is nonlinear in the reponses y_i).

The lasso is an adaptive fitting procedure, and if our final model has k covariates that is different from zero, we would not think that the effective degrees of freedom for the lasso is then k. However, it turns out that it is correct to *count* the number of degrees of freedom by the number of nonzero coefficients.

In ELS Ch 7.6 we also defined the degrees of freedom using the covariance generalization:

$$\mathrm{df}(\hat{\mathbf{y}}) = \frac{\sum_{i=1}^{N} \mathrm{Cov}(\hat{y}_i, y_i)}{\sigma_\varepsilon^2}$$

where the covariance is taken over the reponse variables, while the covariates are kept fixed (this formula was developed in connection to the in-sample prediction error).

It has been shown (HTW refer to this at "somewhat miraculously) that with a fixed penalty parameter λ the number of non-zero coefficients k_λ is an *unbiased estimate* for the degrees of freedom.

This is explained by considering that the lasso does not only select predictors (selecting predictors will give an inflated degrees of freedom) - but also shrinks the coefficients relative to the LS estimates. These two forces kind of cancel out.

HTW (page 19): a general proof is difficult, but for an orthogonal design using the fact that the lasso estimates are soft-thresholded versions f the univariate regression coefficients for the othogonal design.

Finally

- When is ridge preferred to LS? When the LS estimates have high variance and many predictors are truly non-zero.
- Ridge is computationally fast.
- Ridge is not very easy to interpret, because all p predictor are included in the final model.

Neigher ridge or lasso dominates the other in all situations.

Software

We will use the `glmnet` implementation for R:

Exercises

Gauss-Markov theorem

The LS is unbiased with the smallest variance among linear predictors: ELS exercise 3.3a

Variance of ridge compared to LS

Consider a classical linear model with regression parameters β. Let $\hat{\beta}$ be the LS estimator for β and let $\tilde{\beta}$ be the ridge regression estimator for β. Show that $\text{Var}(\hat{\beta}) \geq \text{Var}(\tilde{\beta})$.

Ridge regression

This problem is taken, with permission from Wessel van Wieringen, from a course in High-dimensional data analysis at Vrije University of Amsterdam.

a)

Find the ridge regression solution for the data below for a general value of λ and for the simple linear regression model $Y = \beta_0 + \beta_1 X + \varepsilon$ (only apply the ridge penalty to the slope parameter, not to the intercept). Show that when λ is chosen as 40, the ridge 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 $\mathbf{Y}^T = (Y_1, Y_2, \ldots, Y_8)^T = (35, 40, 36, 38, 40, 43, 45, 43)^T$.

b)

The coefficients β of a linear regression model, $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$, are estimated by $\hat{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$. The associated fitted values then given by $\hat{\mathbf{Y}} = \mathbf{X}\,\hat{\beta} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \mathbf{H}\mathbf{Y}$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$. The matrix $\mathbf{H}$ is a projection matrix and satisfies $\mathbf{H} = \mathbf{H}^2$. Hence, linear regression projects the response $\mathbf{Y}$

Solutions to exercises

Please try yourself first, or take a small peek - and try some more - before fully reading the solutions. Report errors or improvements to Mette.Langaas@ntnu.no.

- Gauss-Markov theorem 3.3a
- Variance of ridge compared to LS: page 11-12 on note by Wessel N. van Wieringen
- Ridge regression

Resources

- Videos in statistics learning with Rob Tibshirani and Daniela Witten, made for the Introduction to statistical learning Springer textbook.
  - Ridge
  - Lasso
  - Selecting tuning parameter
- Video from webinar with Trevor Hastie on glmnet from 2019
- Lecture notes on ridge regression: Welle N. van Wieringen