Compulsory exercise 1: Group 39

TMA4268 Statistical Learning V2021

Alexander J Ohrt, Jim Totland

08 februar, 2021

Problem 1

a)

We assume that $\mathbf{Y}$ is a multivariate normal, which gives the distribution $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$.

$$\begin{aligned}\mathsf{E}(\widetilde{\boldsymbol{\beta}}) &= \mathsf{E}((\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}) = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathsf{E}(\mathbf{Y}) \\ &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathsf{E}(\mathbf{X}\boldsymbol{\beta} + \varepsilon) = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\end{aligned}$$

and

$$\begin{aligned}\mathsf{Cov}(\widetilde{\boldsymbol{\beta}}) &= \mathsf{Cov}((\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}) = ((\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T)\mathsf{Cov}(\mathbf{Y})((\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T)^T \\ &= ((\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T)\sigma^2 I(\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-T}) = \sigma^2((\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-T}) \\ &= \sigma^2((\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}),\end{aligned}$$

where we have used that $(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I})^{-1} = (\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I})^{-T}$ in the last equality (USIKKER PÅ DENNE SISTE! DET BLIR I HVERT FALL HELT RIKTIG Å BEHOLDE -T). In both these equations it is apparent that the moments are equal to those of the OLS estimator when $\lambda = 0$.

b)

The requested moments of $\widetilde{f}(\mathbf{x}_0)$ are

$$\mathsf{E}(\widetilde{f}(\mathbf{x}_0)) = \mathsf{E}(\mathbf{x}_0^T\widetilde{\boldsymbol{\beta}}) = \mathbf{x}_0^T\mathsf{E}(\widetilde{\boldsymbol{\beta}}) = \mathbf{x}_0^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$$

and

$$\begin{aligned}\mathsf{Cov}(\widetilde{f}(\mathbf{x}_0)) &= \mathsf{Cov}(\mathbf{x}_0^T\widetilde{\boldsymbol{\beta}}) = \mathbf{x}_0^T\mathsf{Cov}(\widetilde{\boldsymbol{\beta}})\mathbf{x}_0 \\ &= \sigma^2\mathbf{x}_0^T((\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1})\mathbf{x}_0.\end{aligned}$$

c)

The expected MSE at $\mathbf{x}_0$ is

$$
\begin{aligned}
\mathsf{E}[(y_0 - \widetilde{f}(\mathbf{x}_0))^2] &= [\mathsf{E}(\widetilde{f}(\mathbf{x}_0) - f(\mathbf{x}_0)]^2 + \mathsf{Var}(\widetilde{f}(\mathbf{x}_0)) + \mathsf{Var}(\varepsilon) \\
&= [\mathsf{E}(\widetilde{f}(\mathbf{x}_0)) - \mathsf{E}(f(\mathbf{x}_0))]^2 + \mathsf{Cov}(\widetilde{f}(\mathbf{x}_0)) + \sigma^2 I \\
&= [\mathbf{x}_0^T(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - \mathbf{x}_0^T\boldsymbol{\beta}]^2 + \sigma^2\mathbf{x}_0^T((\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1})\mathbf{x}_0 + \sigma^2 I
\end{aligned}
$$

FOR NOE GRISERI

```
id <- "1X_8OKcoYbng1XvYFDirxjEWr7LtpNr1m"  # google file ID
values <- dget(sprintf("https://docs.google.com/uc?id=%s&export=download", id))
X = values$X
dim(X)
```

```
## [1] 100  81
```

```
x0 = values$x0
dim(x0)
```

```
## [1] 81  1
```

```
beta = values$beta
dim(beta)
```

```
## [1] 81  1
```

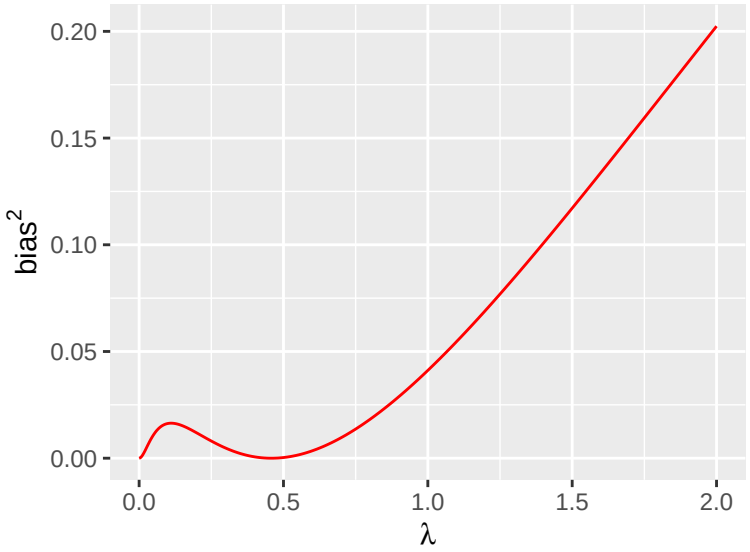
```
sigma = values$sigma
sigma
```

```
## [1] 0.5
```

d)

The expression $\widetilde{\boldsymbol{\beta}}(\lambda) = ((1+\lambda)I)^{-1}\hat{\boldsymbol{\beta}}$ is inserted into `value` below.

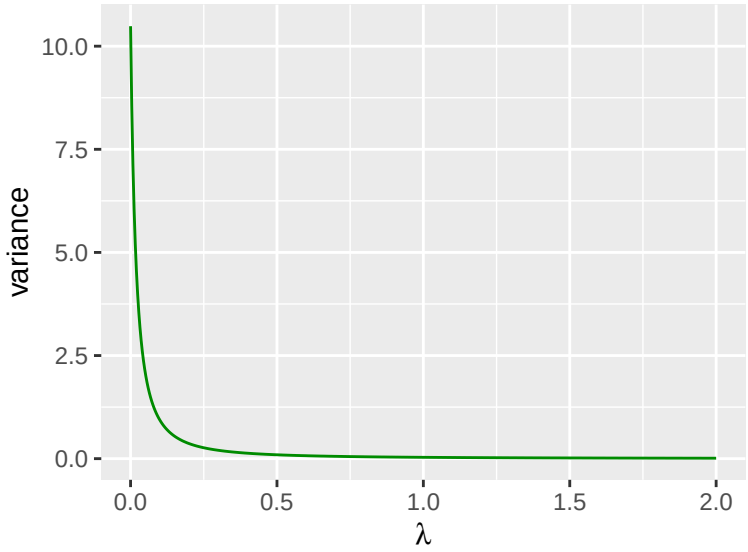
```
library(ggplot2)
bias = function(lambda, X, x0, beta) {
    p = ncol(X)
    lambda * diag(p)
    # value = mean(t(x0) %*% solve((1+lambda)*diag(p)) %*% beta - t(x0) %*% beta )^2
    # value = mean(t(x0) %*% solve(diag(p) + lambda*t(X) %*% X) %*% beta - t(x0) %*%
    # beta )^2
    value = mean(t(x0) %*% solve(t(X) %*% X + lambda * diag(p)) %*% t(X) %*% X %*%
        beta - t(x0) %*% beta)^2
    # Tror dette skal være rett algebraisk, men synes resultatet er merkelig kanskje!
    # Kanskje oppgaven ovenfor burde fullføres først? value = mean(t(x0) %*% (diag(p)
    # + 1/lambda*t(X)%*%X)%*%beta - t(x0) %*% beta)^2

    # All of these give different plots!
    return(value)
}
lambdas = seq(0, 2, length.out = 500)
BIAS = rep(NA, length(lambdas))
for (i in 1:length(lambdas)) BIAS[i] = bias(lambdas[i], X, x0, beta)
dfBias = data.frame(lambdas = lambdas, bias = BIAS)
ggplot(dfBias, aes(x = lambdas, y = bias)) + geom_line(color = "red") + xlab(expression(lambda)) +
    ylab(expression(bias^2))
```



Comments: I think this is wrong. Perhaps a good idea to finish c) first, to get a better expression for the bias also. I think the expression is correct however, despite it not being "forkortet", but this should be done later. However, I think the result is weird.

e)

```
variance = function(lambda, X, x0, sigma) {
    p = ncol(X)
    inv = solve(t(X) %*% X + lambda * diag(p))
    value = sigma * (inv %*% t(X) %*% X %*% inv)
    return(value)
}
lambdas = seq(0, 2, length.out = 500)
VAR = rep(NA, length(lambdas))
for (i in 1:length(lambdas)) VAR[i] = variance(lambdas[i], X, x0, sigma)
dfVar = data.frame(lambdas = lambdas, var = VAR)
ggplot(dfVar, aes(x = lambdas, y = var)) + geom_line(color = "green4") + xlab(expression(lambda)) +
    ylab("variance")
```



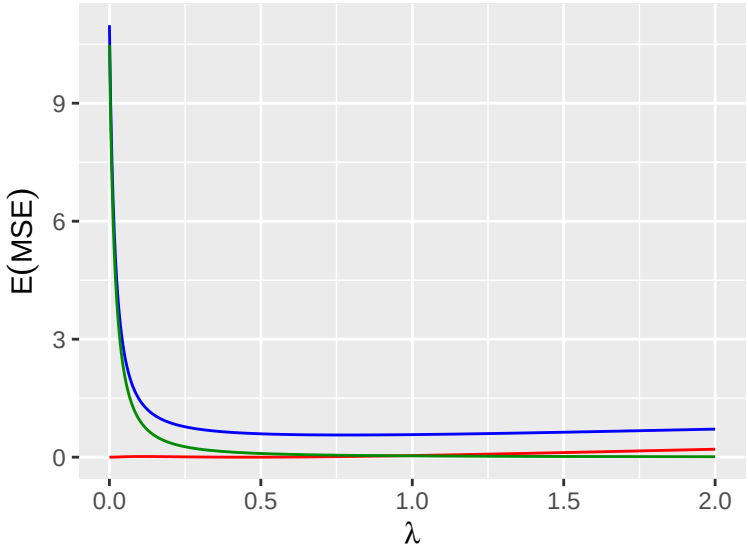
Comments: The variance decreases with lambda. It begins very high (which I do not know if holds with the regular estimator when $\lambda = 0$).

f)

```
exp_mse = BIAS + VAR + sigma
lambdas[which.min(exp_mse)]
```

```
## [1] 0.7815631
```

```
dfAll = data.frame(lambda = lambdas, bias = BIAS, var = VAR, exp_mse = exp_mse)
ggplot(dfAll) + geom_line(aes(x = lambda, y = exp_mse), color = "blue") + geom_line(aes(x = lambda,
    y = bias), color = "red") + geom_line(aes(x = lambda, y = var), color = "green4") +
    xlab(expression(lambda)) + ylab(expression(E(MSE)))
```



Comments: Again, unsure if this is correct.

Problem 2

```
# read file
id <- "1yYlEl5gYY3BEtJ4d7KWaFGIOEweJIn__"  # google file ID
d.corona <- read.csv(sprintf("https://docs.google.com/uc?id=%s&export=download",
    id), header = T)
```

a)

Inspection of the data. Assuming that 0 = False and 1 = True (as usual), which means that a person has not deceased when `deceased` = 0 and vice versa.

```
knitr::kable(table(Deceased = d.corona$deceased))  # Prøvde å sette kolonne-navn, men fikk det ikke til
```

Deceased	Freq
0	1905
1	105

```
names(d.corona)
```

```
## [1] "deceased" "sex"      "age"      "country"
```

```
knitr::kable(table(d.corona$country, d.corona$sex))
```

	female	male
France	60	54
indonesia	30	39
japan	120	174
Korea	879	654

```
knitr::kable(table(d.corona$deceased, d.corona$sex))
```

	female	male
0	1046	859
1	43	62

```
knitr::kable(table(d.corona$country, d.corona$deceased))  # Må hente ut øverste raden herfra, men klart
```

	0	1
France	98	16
indonesia	64	5
japan	283	11
Korea	1460	73

b)

```
# Just in case they are not factors (this should perhaps be deleted later, since
# we could have checked if they were factors earlier).
```

```r
d.corona$sex = factor(d.corona$sex)
d.corona$country = factor(d.corona$country)
lm.fit <- lm(deceased ~ ., data = d.corona)  # perhaps a linear model is not the correct model to use?
summary(lm.fit)
```

```
##
## Call:
## lm(formula = deceased ~ ., data = d.corona)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.20383 -0.07105 -0.04495 -0.02110  1.03018
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.043862   0.025229   1.739 0.082263 .
## sexmale            0.030815   0.009902   3.112 0.001884 **
## age                0.001305   0.000218   5.984 2.57e-09 ***
## countryindonesia -0.053478   0.033584  -1.592 0.111455
## countryjapan     -0.097525   0.024269  -4.018 6.08e-05 ***
## countryKorea     -0.071966   0.021542  -3.341 0.000851 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2193 on 2004 degrees of freedom
## Multiple R-squared:  0.03144,    Adjusted R-squared:  0.02902
## F-statistic: 13.01 on 5 and 2004 DF,  p-value: 1.755e-12
```

(i) The probability to die of covid for a male age 75 living in Korea can be predicted from the model. The prediction is ...

(ii) The p-value for `sexmale` is relatively small, but we would not say that there is clear evidence that males have higher probability to die than females. The p-value could be low by chance also, and since it is not amazingly small, we do not think it is appropriate evidence of the question.

(iii) The p-value of `countryjapan` is the smallest of the country-coefficients, which shows that it could potentially be important as a predictor. In the least, it does not exclude this possibility.

(iv) Quantify the odds

c)

d)