UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	STK9030 — Statistical Learning: Advanced Regression and Classification
Day of examination:	Friday 11th of December
Examination hours:	14.30 – 18.30
This problem set consists of 4 pages.	
Appendices:	None
Permitted aids:	Approved calculator

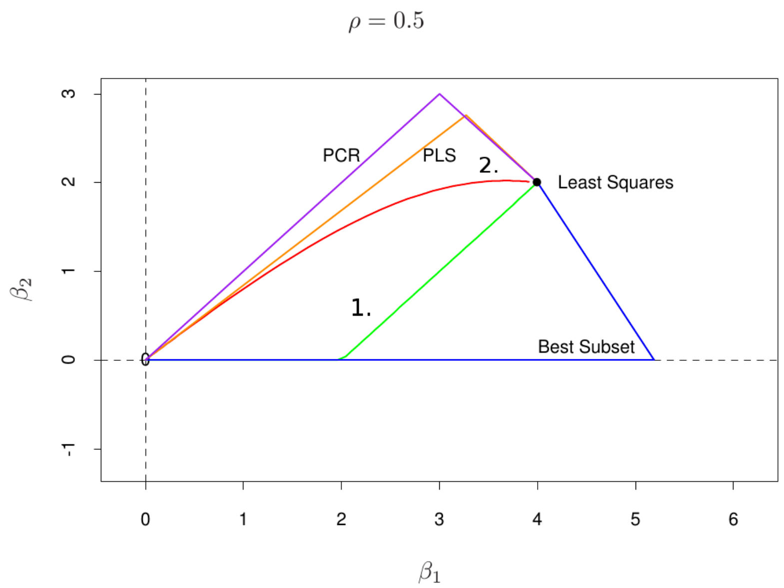
Please make sure that your copy of the problem set is complete before you attempt to answer anything.

Problem 1 Penalized regression

A linear regression model with no intercept and two input variables is given as

$$f(x) = x_1\beta_1 + x_2\beta_2 + \epsilon.$$

The figure below shows profiles of different coefficient estimates as the tuning parameter varies between 0 and ∞. Five linear regression methods are shown: partial least squares (PLS), principal component regression (PCR) and best subset regression and two methods from the curriculum. The ordinary least squares (OLS) solution is shown by the point $(\hat{\beta}_1^{OLS}, \hat{\beta}_2^{OLS}) = (4, 2)$ as OLS has no tuning parameter.



(Continued on page 2.)

a

The unmarked paths are given by two penalized regression methods with different penalties. Specify the methods 1 and 2 and give the penalized residual sum of square (PRSS) for each.

b

Explain shortly in terms of the penalties why their paths are different. Which characteristic does method 1 exhibit?

Problem 2 Ridge regression

This problem will explore how ridge regression handles correlated inputs.

Consider two correlated inputs $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ with expectation zero and covariance matrix $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ and an output Y given by a linear regression model with no intercept

$$Y = X^T \boldsymbol{\beta} + \epsilon, \text{ where } \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}.$$

For N observations and p inputs, the ridge regression solution is given by the data matrix $\boldsymbol{X}$ and response vector $\boldsymbol{y}$ as

$$\hat{\boldsymbol{\beta}}^{ridge} = \begin{bmatrix} \hat{\beta}_1^{ridge} \\ \hat{\beta}_2^{ridge} \end{bmatrix} = (\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

a

Show that

$$\hat{\boldsymbol{\beta}}^{ridge} = \boldsymbol{A} \hat{\boldsymbol{\beta}}^{OLS},$$

where $\boldsymbol{A}$ is a matrix depending on $\boldsymbol{X}$ and λ, meaning that the ridge solution is linear combination of the OLS solution.

b

For a large number of observations N, one can simplify calculations by using the following approximation

$$\boldsymbol{X}^T \boldsymbol{X} \simeq N\Sigma. \tag{1}$$

Find approximate expressions of the ridge coefficients $\hat{\beta}_1^{ridge}$ and $\hat{\beta}_2^{ridge}$ for large N, as weighted sums of $\hat{\beta}_1^{OLS}$ and $\hat{\beta}_2^{OLS}$ where the weights depend on ρ, λ and N.

In the case of $\rho > 0$, how will ridge regression shrink the regression coefficients?

Note: the inverse of a 2×2 matrix is given

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

(Continued on page 3.)

c

The elastic net method combines the lasso and ridge penalty:

$$PRSS^{elastic}_{\lambda,\alpha}(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda(1-\alpha)\sum_{j=1}^{p}|\beta_j| + \lambda\alpha\sum_{j=1}^{p}\beta_j^2.$$

Explain how this combination improves on the separate methods.

Problem 3 Crossvalidation

Describe in detail how k-fold crossvalidation is used to select a tuning parameter, for instance λ in lasso regression.

Problem 4 Principal components

The singular value decomposition of a design matrix is given $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$, where $\boldsymbol{V}^T\boldsymbol{V} = I$ and $\boldsymbol{U}^T\boldsymbol{U} = I$. The columns of $\boldsymbol{U}$, $\boldsymbol{u}_i$, are referred to as the principal component directions.

a

Express the ridge and PCR predictions $\hat{\boldsymbol{y}}^{ridge}_{\lambda} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{ridge}_{\lambda}$ and $\hat{\boldsymbol{y}}^{PCR}_{m} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{PCR}_{m}$ in terms of $\boldsymbol{y}$, $\boldsymbol{u}_i$, $\boldsymbol{d}_i$ and their respective tuning parameters.

b

Explain how the ridge and PCR predictions shrink the quantity $\boldsymbol{u}_i\boldsymbol{u}_i^T\boldsymbol{y}$ differently. Sketch for instance a figure showing the shrinkage factor as a function of the principal component index (the column number i).

Problem 5 Boosting and bagging

a

Describe shortly the main concept behind

i) boosting

ii) bagging

b

In which way does the random forest method aim to improve on standard tree bagging? Which step of the tree bagging algorithm is modified to achieve this?

(Continued on page 4.)

c AdaBoost

The algorithm below shows the boosting classification algorithm AdaBoost.M1, which considers responses $Y \in \{-1, 1\}$ with an exponential loss function and a general base classifier $G_m(x)$.

Algorithm 10.1 *AdaBoost.M1.*

---

1. Initialize the observation weights $w_i = 1/N,\ i = 1, 2, \ldots, N$.
2. For $m = 1$ to M:
   (a)
   (b) Compute
   $$\text{err}_m = \frac{\sum_{i=1}^{N} w_i I(y_i \neq G_m(x_i))}{\sum_{i=1}^{N} w_i}.$$
   (c) Compute $\alpha_m = \log((1 - \text{err}_m)/\text{err}_m)$.
   (d) Set $w_i \leftarrow w_i \cdot \exp[\qquad\qquad\qquad\qquad]$, $i = 1, 2, \ldots, N$.
3. Output $G(x) = \text{sign}\left[\qquad\qquad\qquad\right]$.

Give a description of the missing step in line $2(a)$ and give the two missing expressions in line $2(d)$ and 3.

SLUTT