UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Examination in:	STK4030/9030 — Modern data analysis
Day of examination:	Friday, December 13th, 2013.
Examination hours:	14.30 – 18.30.
This problem set consists of 6 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

Consider a regression situation where $Y = f(\boldsymbol{X}) + \varepsilon$. We want to predict Y given $\boldsymbol{X} = \boldsymbol{x}_0$.

(a) We define the *expected prediction error* by

$$\text{EPE}(\boldsymbol{x}_0) = E[(Y - \widehat{Y}(\boldsymbol{x}_0))^2 | \boldsymbol{X} = \boldsymbol{x}_0]$$

Find an expression that relates EPE to bias and variance. Discuss this result.

Assume a loss function $L(Y, \widehat{Y}(\boldsymbol{X})) = (Y - \widehat{Y}(\boldsymbol{X}))^2$. Find the optimal prediction in this case.

Assume now that $f(\boldsymbol{X})$ is unknown but we have for our disposal a training sample $\{(y_i, \boldsymbol{x}_i), i = 1, ..., N\}$. In the following we will use the notation $\boldsymbol{y} = (y_1, ..., y_n)^T$ and $\boldsymbol{X}$ the matrix with $\boldsymbol{x}_i^T$ at the ith row.

(b) Consider two possible predictors

$$\hat{Y} = \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}} \qquad \text{where } \hat{\boldsymbol{\beta}} = [\boldsymbol{X}^T \boldsymbol{X}]^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

$$\hat{Y} = \frac{1}{k} \sum_{\boldsymbol{x}_i \in N_k(\boldsymbol{x}_0)} y_i$$

(Continued on page 2.)

where $N_k(\boldsymbol{x})$ is the neighborhood of $\boldsymbol{x}$ defined by the k closest points $\boldsymbol{x}_i$ in the training sample. Discuss benefits and weaknesses of each of these two predictors by relating them to the previous results.

Problem 2

Consider now a linear regression model

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \varepsilon$$

Assume $\boldsymbol{\beta} = (\beta_0, ..., \beta_p)$ is estimated by the criterion

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \qquad (*)$$

(a) What is this method called? Discuss this method compared to ordinary least squares.

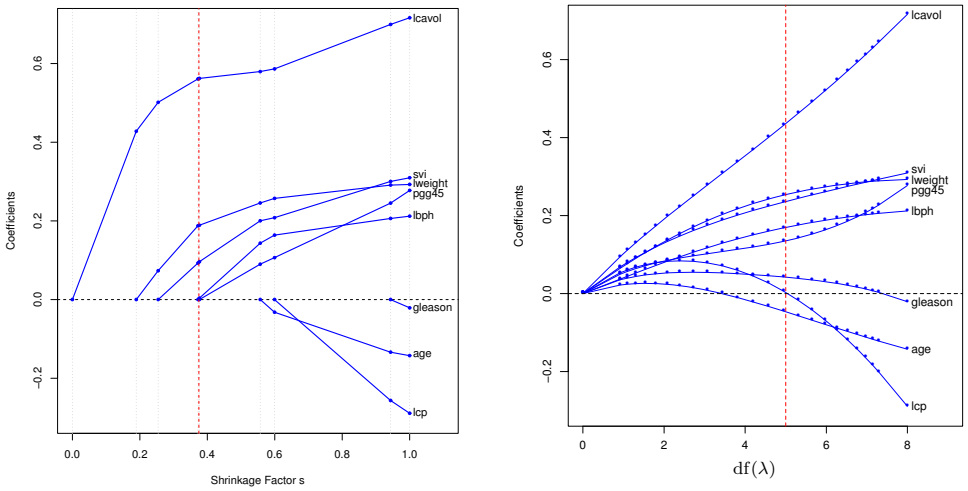
(b) Find an explicit expression for $\hat{\boldsymbol{\beta}}$.

An alternative way of estimating $\boldsymbol{\beta}$ is through the criterion

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} \qquad (**)$$

(c) What is this method called? Discuss this method in relation to both ordinary least squares and the method above.

Below are two plots showing the estimates of $\boldsymbol{\beta}$ on the prostate cancer example from the textbook. Relate the two plots to the two methods discussed above. Also explain how the least squares estimates can be read from the plot(s).



(Continued on page 3.)

(d) In both cases, the penalty variable λ needs to be specified. Discuss methods for doing that.

Problem 3

Consider now a classification problem where the response now is within a discrete set $\mathcal{G} = \{1, ..., K\}$. For simplicity we will concentrate on the case $K = 2$

(a) Assume

$$\Pr(G = 1|\boldsymbol{X} = \boldsymbol{x}) = \frac{\exp(\beta_0 + \boldsymbol{\beta}^T\boldsymbol{x})}{1 + \exp(\beta_0 + \boldsymbol{\beta}^T\boldsymbol{x})}$$

Explain why we view this as a *linear* method for classification.

(b) Discuss how estimation of β_0 and $\boldsymbol{\beta}$ can be performed. In particular, describe how the ideas behind the criterions $(*)$ and $(**)$ in Problem 2 can be applied also in this case.

Problem 4

In this exercise, you will first be introduced to a problem about traffic related air pollution, and then you will be asked to interpret a GAM plot.
Particles between 2.5 and 10 micrometers in size are called coarse particles. At or beside a road in a Norwegian city, there will typically be a lot of road dust containing many coarse particles, especially in the winter when many cars are fitted with studded tyres (piggdekk). These particles are typically whirled into the air by the cars. In addition to the road dust which is re-suspended into the air, the exhaust from the vehicles also gives direct emissions of coarse particles (in addition to emissions of smaller particles and gases).
A large data set has been analysed to give a description of how the concentration of coarse particles are related to the traffic volume and meteorological conditions. This data set consists of 70 000 hourly values of the concentration of coarse particles and corresponding explanatory variables in the period 2001-2009 at Kirkeveien in Oslo. The speed limit at Kirkeveien is 50 km per hour. The variables used in this analysis includes

- y - the (natural) logarithm of the concentration of coarse particles in the air
- x_1 = the number of light vehicles (shorter or equal to 5.5 m) per hour = "trafikkLette"
- x_2 = the number of heavy vehicles (longer than 5.5 m) per hour = "trafikkTunge"

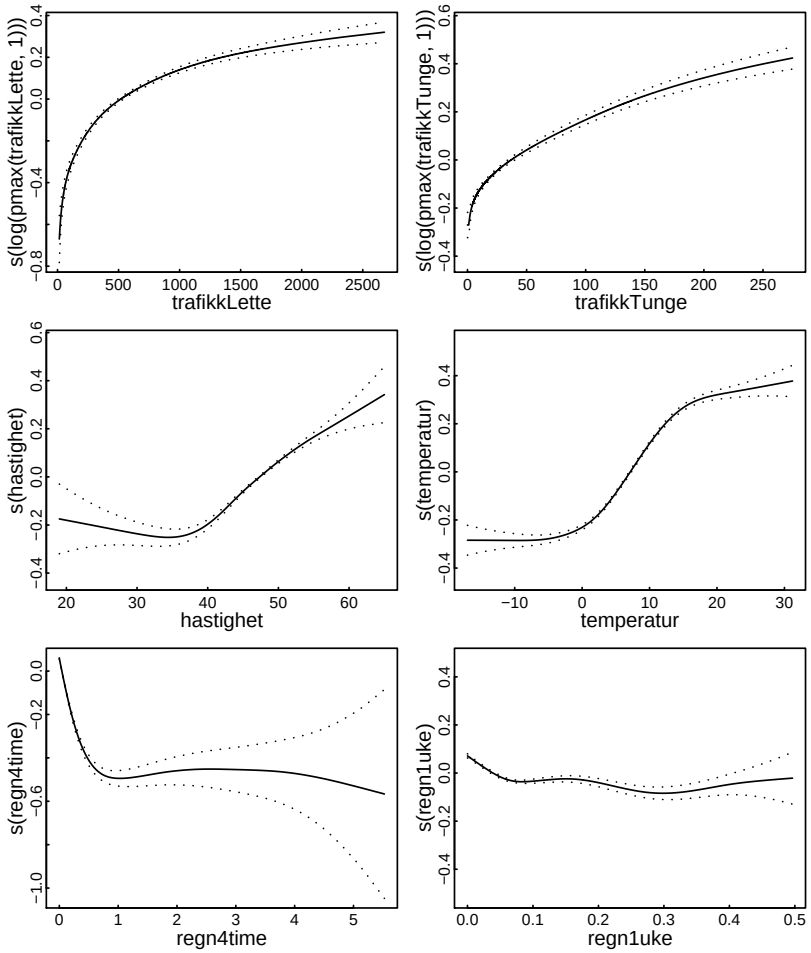
(Continued on page 4.)

- x_3 = average velocity of the vehicles = "hastighet"
- x_4 = temperature in degrees Celsius = "temperatur"
- x_5 = accumulated precipitation (in mm) last 4 hours = "regn4time"
- x_6 = accumulated precipitation (in mm) last week = "regn1uke"
- other explanatory variables that you can ignore in this exercise

The following generalised additive model (GAM) has been fitted to the data

$$y = s_1(x_1) + s_2(x_2) + s_3(x_3) + s_4(x_4) + s_5(x_5) + s_6(x_6) + \ldots + \varepsilon.$$

Here, $+\ldots$ mean that also some other explanatory variables are included in the model. The figure below shows the GAM plot for x_1-x_6.



(a) Give a short interpretation of the estimated effect of each of the six explanatory variables, i.e. describe how they may affect the concentration of coarse particles, and discuss whether the results are reasonable i light of your understanding of the physical process of this type of air pollution.

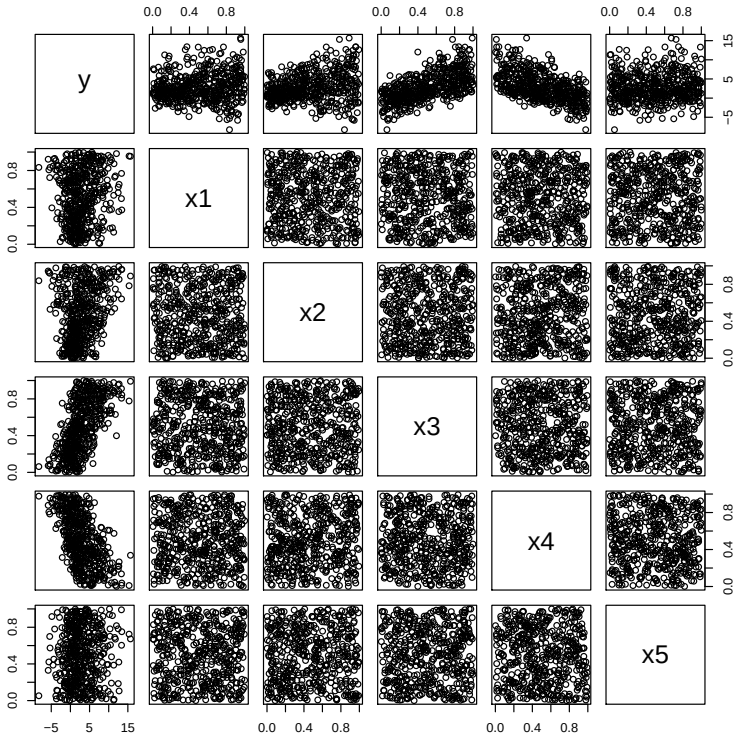
(Continued on page 5.)

(b) Until a few years ago, the speed limit at some of the main roads in Oslo was lowered from 80 km/hour to 60 km/hour during the winter months. Use the GAM plot above to quantify what effect this intervention might have had on the concentration on coarse particles near these roads.

Problem 5

(a) Describe what a projection pursuit regression model is, both by words and by writing the model as a formula. What are the tuning parameters in this model?

(b) Consider now a specific regression problem with 500 observations of a response y and five explanatory variables x_1-x_5. The Figure below shows scatter plots for all pairs of variables. The standard deviation of y is 3.76.



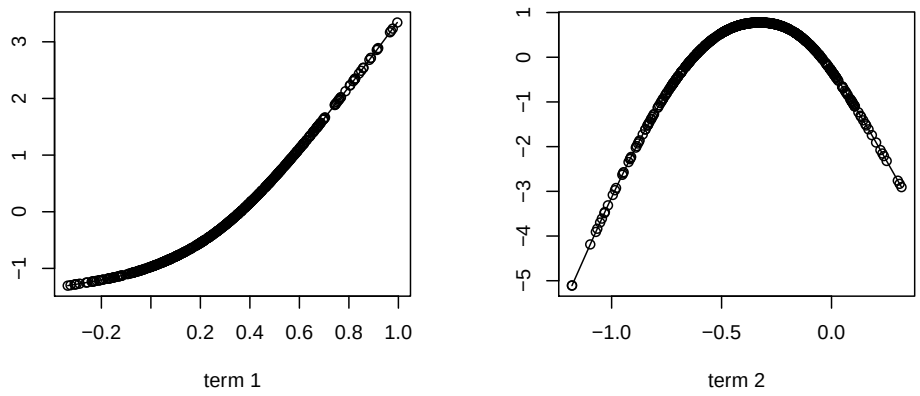
A projection pursuit regression model has been fitted to these data, and the tuning parameters has been chosen by 10-fold cross validation. The prediction root mean squared error from the cross validation with optimal tuning parameters is 0.98, whereas the residual standard error for the final fitted model is 0.93.

Below, you find an exerpt from the computer output from the function `ppr` in the `MASS` library in R, and the figure below shows a plot of the fitted non-linear functions. Give a short interpretation of the various parts of the model.

(Continued on page 6.)

```
Projection direction vectors:
    term 1       term 2
Xx1  0.205750216 -0.424701344
Xx2  0.389862442 -0.306911516
Xx3  0.637864019  0.584250950
Xx4 -0.631484540 -0.619707250
Xx5  0.005558274  0.006916854

Equivalent df for ridge terms:
term 1 term 2
  4.83   5.46
```



Does the model explain the relationship between y and the x's reasonably well?

Do you think the following generalised additive model would be able to model the relationship between y and the x-es better or worse than the projection pursuit regression model?

$$y = s_1(x_1) + s_2(x_2) + s_3(x_3) + s_4(x_4) + s_5(x_5) + \varepsilon.$$

END