TMA4300 Computer Intensive Statistical Methods
Exercise 3, Spring 2014

Note: The solution to ALL exercises must be handed in no later than **April 28th 2014**.
Hint: In almost all exercises you will need to use the R-function `sample`.

Problem A: Classification and cross validation

In this exercise we will use the data set `kyphosis` (available in R). Load the library `rpart` and the data set by `data(kyphosis)`. The first column, `Kyphosis`, is an indicator for presence and absence of a disease, and `Age`, `Number` and `Start` are predictor variables. Read the online help, available via `?kyphosis`, on the data set before analysing it.

1. Fit a linear and a quadratic discriminant model to the kyphosis data set, and estimate the misclassification rates by 10-fold cross validation. (Hint: You can use the R-function `lda` and `qda` available in the library `MASS`)

2. Construct a k-nearest neighbour classifier to the kyphosis data set. Find the best value for k by 10-fold cross validation. (Hint: You can use the R-function `knn` available in the library `class`)

Problem B: Comparing $AR(2)$ parameter estimators using resampling of residuals

The data files and pre-programmed R-code can be downloaded from the course webpage. Look in the `probBhelp.R`-file and read the documentation to see how the code works. Load the code and data into `R` with

```
source("probBhelp.R")
source("probBdata.R")
```

In this exercise you should analyse the data in `data3A$x`, which contains a sequence of length $T = 100$ of a non-Gaussian time-series, and compare two different parameter estimators.
We consider an AR(2) model which is specified by the relation

$$x_t = \beta_1 x_{t-1} + \beta_2 x_{t-2} + \epsilon_t,$$

where e_t are iid random variable with zero mean and constant variance.
The least sum of squared residuals (LS) and least sum of absolute residuals (LA) are obtained by minimising the following loss functions with respect to $\boldsymbol{\beta}$:

$$Q_{LS}(\boldsymbol{x}) = \sum_{t=3}^{T} (x_t - \beta_1 x_{t-1} - \beta_2 x_{t-2})^2$$

$$Q_{LA}(\boldsymbol{x}) = \sum_{t=3}^{T} |x_t - \beta_1 x_{t-1} - \beta_2 x_{t-2}|$$

Denote the minimisers by $\widehat{\boldsymbol{\beta}}_{LS}$ and $\widehat{\boldsymbol{\beta}}_{LA}$ (calculated by `ARp.beta.est`), and define the estimated innovations to be $\hat{e}_t = x_t - \hat{\beta}_1 x_{t-1} - \hat{\beta}_2 x_{t-2}$ for $t = 3, \ldots, T$, and let $\bar{e}$ be the mean of these. The $\hat{e}_t$ can be re-centered to have mean zero by defining $\hat{\epsilon}_t = \hat{e}_t - \bar{e}$. (Results for $\hat{\epsilon}_t$ obtained by LS and LA can be calculated with `ARp.resid`).

1. Use the residual resampling bootstrap method to evaluate the relative performance of the two parameter estimators. Specifically, estimate the variance and bias of the two estimators.

   You may use `ARp.filter` as a helper function in your resampling code. Use at least $B = 1500$ bootstrap samples, each as long as the original data sequence ($T = 100$). To do a resampling, you first need to initialise values for x_1 and x_2 by picking a random consecutive subsequence from the data.

   The LS estimator is optimal for Gaussian AR(p) processes. Is it also optimal for this problem?

2. Compute a 95% prediction interval for x_{101} based on both estimators. That means using the corresponding parameter estimates obtained in part 1), predict for each bootstrap iteration a value for x_{101}. From these B predictions derive a 95% quantile-based confidence interval.

Problem C: Permutation test

Bilirubin (see `http://en.wikipedia.org/wiki/Bilirubin`) is a breakdown product of haemoglobin, which is a principal component of red blood cells. If the liver has suffered degeneration, if the decomposition of haemoglobin is elevated, or if the gall bladder has been destroyed, large amounts of bilirubin can accumulate in the blood, leading to jaundice. The following data (taken from Jørgensen (1993)) contain measurements of the concentration of bilirubin (mg/dL) in blood samples taken from three young men.

Individual	Concentration (mg/dL)										
1	0.14	0.20	0.23	0.27	0.27	0.34	0.41	0.41	0.55	0.61	0.66
2	0.20	0.27	0.32	0.34	0.34	0.38	0.41	0.41	0.48	0.55	
3	0.32	0.41	0.41	0.55	0.55	0.62	0.71	0.91			

We will use the F-statistic to perform a permutation test.

1. Download the data file `bilirubin.txt` from the course webpage and read it into R using

   ```
   bilirubin <- read.table("bilirubin.txt",header=T)}
   ```

   The first column, labelled `meas`, contains the concentrations (mg/dL) as shown in the table. The second column, `pers`, is an indicator for the individual.

   Use a boxplot to inspect the logarithms of the concentrations for each individual.

2. Use the function `lm` in `R` to fit the regression model

$$\log Y_{ij} = \beta_i + \epsilon_{ij}, \qquad \text{with } i = 1, 2, 3 \text{ and } j = 1, \ldots, n_i \tag{1}$$

   where $n_1 = 11$, $n_2 = 10$ and $n_3 = 8$, and $\epsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. Use the F-test to test the hypothesis that $\beta_1 = \beta_2 = \beta_3$ and save the value of the F-statistic as `Fval`. Is the hypothesis accepted? (Hint: The value of the F-statistic is contained in the default output of `lm`)

3. Write a function `permTest()` which generates a permutation of the data between the three individuals, consequently fits the model given in (1) and finally returns the value of the F-statistic for testing $\beta_1 = \beta_2 = \beta_3$.

4. Perform a permutation test using the function `permTest` to generate a sample of size 999 for the F-statistic. Compute the p-value for `Fval` in this sample. What do you observe?

Problem D: EM-algorithm

Consider the following two-way table of y_{ij} for $i = 1, 2$ and $j = 1, 2, 3$ with one missing cell y_{22}:

5	8	7
10	•	12

As in the lecture consider a linear model

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

where $\sum_i \alpha_i = \sum_j \beta_j = 0$ and $\epsilon_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$.

1. Write an R-function to impute the missing value using the EM-algorithm, and describe the single steps you perform. Stop the algorithm if the absolute change in y_{22} is smaller than $1e-5$.

2. Plot the estimated parameters μ, α_1, α_2, β_1, β_2, β_3, as well as the value of y_{22} as a function of the EM iteration number.

Literature

Jørgensen, B. (1993). The Theory of Linear Models. Chapman and Hall

Oral presentations

Date	Exercise	Team
20.02.2014	1: Problem A1 and A2	Marius Møller Rokstad
	1: Problem A3	Ilmo Räisänen
	1: Problem B	Lars Kristian Steffensen, Shipra Sachdeva
	1: Problem C1 and C2	Tygve Bertelsen Wiig
27.02.2014	1: Problem C3	Henrik Vikøren, Edvard Hove
	1: Problem C4 and D1	Elise Landsem, Margrethe Kvale Loe
	1: Problem D2	Mateusz Samiec
20.03.2014	2: 1 a, b	Tore Bredre
	2: 1c (GI)	Andrea Casati
	2: 1c (BL)	— open discussion —
	2: 2 (GI)	Ekaterina Fedorova, Beate Sildnes
27.03.2014	2: 2 (BL)	Pål Christie Ryalen
	2: 3	Odd Eirik Farestveit, Susanne Kjølen
	2: 4,5	Marius Fagerland, Tobias Bjormyr
	2: 6,7	Torgeir Rimstad, Kristoffer Berg
10.04.2014	3: Problem A	Brandon Bergerud
	3: Problem B	Kristoffer Kofoed Rødvei, Sverre Thommesen
	3: Problem C	Kristin M. Drahus
	3: Problem D	James Korley Attuquaye, Mireia Duaso