TMA4300 Computer Intensive Statistical Methods
Exercise 3, Spring 2012

Note: The solution of problems A, B and C should be handed in no later than **April 30th 2011**.

The data files and pre-programmed R-code can be downloaded from the course webpage. Look in the `prob3help.R`-file to read the documentation, and see how the code works. In addition, you will need to use the function `sample` in your own Bootstrap implementations. Load the code and data into R with

```
source("prob3help.R")
source("prob3data.R")
```

Problem A: Comparing $AR(p)$ parameter estimators using resampling of residuals

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.

Given some initial values $\mathbf{x}_0 = \{x_{1-p}, x_{1-p+1}, \ldots, x_{-1}, x_0\}$

$$
\begin{aligned}
&e_t \sim \text{independent, identically distributed, mean } 0\\
&x_t = \beta_1 x_{t-1} + \ldots + \beta_p x_{t-p} + e_t, \quad t = 1, \ldots, T
\end{aligned}
$$

The relationship between the observed quantities and the residuals can be written in matrix form:

$$
\begin{bmatrix} x_{p+1} \\ \vdots \\ x_T \end{bmatrix} = \mathbf{y} = \mathbf{C}\beta + \mathbf{e} = \begin{bmatrix} x_p & \cdots & x_1 \\ \vdots & & \vdots \\ x_{T-1} & \cdots & x_{T-p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_{p+1} \\ \vdots \\ e_T \end{bmatrix}
$$

The least sum of squared squared residuals (LS) and least sum of absolute residuals (LA) are obtained by minimising the following loss functions with respect to β:

$$
\begin{aligned}
Q_{LS}(\mathbf{x}) &= \sum_{t=p+1}^{T} \left(x_t - \sum_{k=1}^{p} x_{t-k}\beta_k \right)^2 = \|\mathbf{y} - \mathbf{C}\beta\|_2^2\\
Q_{LA}(\mathbf{x}) &= \sum_{t=p+1}^{T} \left| x_t - \sum_{k=1}^{p} x_{t-k}\beta_k \right| = \|\mathbf{y} - \mathbf{C}\beta\|_1
\end{aligned}
$$

Denote the minimisers by $\widehat{\beta}_{LS}$ and $\widehat{\beta}_{LA}$ (calculated by `ARp.beta.est`), and define the observed residuals as $\widehat{e} = \mathbf{y} - \mathbf{C}\widehat{\beta}$ (different values for LS and LA, can be calculated with `ARp.resid`). You can assume that $p = 2$ is known.

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 resample the `x0` sequence (of length p) by picking a random subsequence from the data.

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

2. Compute prediction intervals for x_{101}, based on Bootstrap, one for each parameter estimator.

Problem B: Permutation test for two samples

Here, you will test if the data in `data3B$y` `data3B$z` have the same distribution.
The simple model for independent data from two sources that you should use is the following:

$$
\begin{aligned}
y_i &\sim F_1, \quad i = 1, \ldots, m \\
z_j &\sim F_2, \quad j = 1, \ldots, n \\
\mathbf{x} &= (\mathbf{y}, \mathbf{z}) = (y_1, \ldots, y_m, z_1, \ldots, z_n)
\end{aligned}
$$

The permutation method for hypothesis testing is based on resampling under the null hypothesis $H_0 : F_1 = F_2$, by permuting the order of the original data (use `sample(x, ..., replace=FALSE)`) to generate B Bootstrap samples $\mathbf{x}^*$ valid given that the null hypothesis is true. The p-value for a test based on a test quantity $T(\mathbf{x})$ can then be estimated as $\#\{T(\mathbf{x}^*) \geq T(\mathbf{x})\}/B$. The null hypothesis is rejected if the p-value is smaller than a given threshold (typically 0.05 or 0.01)

1. Test the hypothesis

   $$H_0 : F_1 = F_2$$

   against

   $$H_1 : F_1 \neq F_2$$

   using the test quantity $T = |\overline{y} - \overline{z}|$, using the permutation method to compute an estimate of the p-value for the test.

2. The test only tests for differences that can be detected by the test quantity. Calculate the p-value based on the alternative test quantity $T = \left| \frac{\left(\frac{1}{m}\sum_{i=1}^m y_i\right)^2}{\frac{1}{m}\sum_{i=1}^m y_i^2} - \frac{\left(\frac{1}{n}\sum_{j=1}^n z_j\right)^2}{\frac{1}{n}\sum_{j=1}^n z_j^2} \right|$ and compare the result to the previous p-value.

Problem C: Estimating prediction error using cross-validation

The available training data in `data3C$x` (the same data as in problem B, formatted differently) contains pairs of group indices g and measured values y. The assumed model takes the following form:

$$
\begin{aligned}
g &\sim (\pi_1, \pi_2), \quad \pi_1 + \pi_2 = 1, \, \pi_1, \pi_2 \geq 0 \\
(y|g = k) &= F_k \\
\mathbf{x} &= ((g_1, y_1), \ldots, (g_n, y_n))
\end{aligned}
$$

1. Show that an optimal Bayesian classifier based on an assumption of Exponential models

   $$p(y|g = 1) = \lambda_1 \exp(-\lambda_1 y), \, y > 0, \quad \text{and} \quad p(y|g = 2) = \lambda_2 \exp(-\lambda_2 y), \, y > 0,$$

   is given by

   $$\widehat{g} = \begin{cases} 1, & \text{if } (\lambda_1 - \lambda_2)y < \log \pi_1\lambda_1 - \log \pi_2\lambda_2, \\ 2, & \text{if } (\lambda_1 - \lambda_2)y > \log \pi_1\lambda_1 - \log \pi_2\lambda_2. \end{cases}$$

2. Calculate the estimate $\widehat{\theta}(\mathbf{x}) = (\widehat{\pi_1}, \widehat{\pi_2}, \widehat{\lambda_1}, \widehat{\lambda_2})$ of the parameters $\theta = (\pi_1, \pi_2, \lambda_1, \lambda_2)$ and write an R-function that calculates the optimal classifier from the step above (the arguments should be the parameters and the y-values that should be classified).

3. If the data are not Exponential, the classifier may not be optimal, and directly analysing its properties is difficult. Instead, use cross-validation to estimate the expected classification error, without assuming Exponential data. Divide the data into $K \geq 10$ random but disjoint subgroups $\mathbf{x}^k$ (Hint: use `sample` to calculate a random permutation index vector).

   Let $\widehat{g}_i^{-k(i)}$ denote the classifier of y_i based on the parameter estimate from the data in the subgroups not containing i. The estimated classification error becomes

$$\widehat{PE}_{CV} = \frac{1}{K} \sum_{k=1}^{K} PE(\mathbf{x}^k;\, \theta = \widehat{\theta}(\mathbf{x}^{-k})) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\widehat{g}_i^{-k(i)} \neq g_i)$$

   Write an R-function that calculates $\widehat{PE}_{CV}$.

4. Calculate the cross-validation predicion error and compare with the naive error estimate

$$\widehat{PE}_0 = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\widehat{g}_i \neq g_i)$$

   where $\widehat{g}_i$ is calculated based on $\widehat{\theta}(\mathbf{x})$. (The difference $\widehat{PE}_{CV} - \widehat{PE}_0$ is an estimate of the *optimism* of the estimator $\widehat{PE}_0$.)