Advanced Statistical Computing 2012 - EXAM part 1

Problem 1.

a Make an informative perspective plot of the density f given by, for ϕ the standard normal density,

$$f(x,y) = \tfrac{1}{3}\phi(x)\phi(y) + \tfrac{1}{2}\phi(x-4)\phi(y-3) + \tfrac{1}{6}\phi(x)\phi(y-3).$$

b Write a program to simulate a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from this density.

c Compute by simulation the correlation coefficient $\rho(X, Y)$ of two variables that are distributed according to this density.

Let r_{10} be the sample correlation coefficient of $(X_1, Y_1), \ldots, (X_{10}, Y_{10})$. Suppose a null hypothesis is that these variables are a sample of size 10 from the density $(x, y) \mapsto \phi(x)\phi(y)$ (i.e. the variables are i.i.d. standard normal).

d Compute a critical value c such that the test "reject H_0 if $r_{10} > c$" has size (=power under H_0) 0.05.

e Compute the power of this test at the density f.

Problem 2. To estimate the age at which a person first experiences migraine questionaires were distributed over a random sample from a population, posing the two questions:

- What age do you have now?
- Did you ever experience a migraine?

Thus data $(C_1, \Delta_1), \ldots, (C_n, \Delta_n)$ on a random sample of n persons were obtained, with C_i the age of a person and

$$\Delta_i = \begin{cases} 1, & \text{if } T_i \le C_i, \\ 0, & \text{if } T_i > C_i. \end{cases},$$

where T_i is the age at first migraine. The variables $T_1, \ldots, T_n$ were not observed.

Assume that $T_1, \ldots, T_n$ are exponentially distributed with rate λ, i.e. they have probability density function $f(t) = \lambda e^{-\lambda t}$, for $t > 0$.

(Work out the answers to questions a and c by pen and paper, and hand in the paper.)

a Show that the EM-algorithm for calculating the MLE of λ based on the observed data, with $(C_1, T_1), \ldots, (C_n, T_n)$ the full data, leads to the iterations

$$\lambda_{new} = \lambda_{old}\left(\frac{1}{n}\sum_{i=1}^{n}\Big(\Delta_i \frac{G(C_i, 2, \lambda_{old})}{G(C_i, 1, \lambda_{old})} + (1-\Delta_i)\frac{1 - G(C_i, 2, \lambda_{old})}{1 - G(C_i, 1, \lambda_{old})}\Big)\right)^{-1},$$

for $G(\cdot, k, \lambda)$ the cumulative distribution function of the $\Gamma(k, \lambda)$ distribution:

$$G(c, 1, \lambda) = \int_0^c \lambda e^{-\lambda t}\, dt, \qquad G(c, 2, \lambda) = \int_0^c \lambda^2 t e^{-\lambda t}\, dt.$$

(This function is available in `R` as `pgamma(c,k,λ)`.)

b Implement the EM-algorithm, and use it to calculate the MLE for λ for the data in the file `migraine.txt`. (N.B. Be careful with your choice of starting point; a bad value will easily lead to division by zero.)

c Show that the likelihood of the observed data is given by

$$\prod_{i=1}^{n}(1 - e^{-\lambda C_i})^{\Delta_i} e^{-\lambda C_i(1-\Delta_i)} p(C_1, \ldots, C_n).$$

(The last term is the density of $(C_1, \ldots, C_n)$, which is assumed not to depend on λ, and may also be omitted.)

d Calculate the MLE for λ using a Newton-Rhapson type optimizer on the log likelihood for the observed data.

e Use the output of the algorithm to give an estimate of the standard error of the MLE.

Advanced Statistical Computing 2012 - EXAM part 2

Problem 3. The observations $X_1, \ldots, X_{10}$ are a random sample from a shifted t-distribution with mean θ and 4 degrees of freedom (i.e. each X_i is the sum of θ and a standard t_4-variable).

a Compute by simulation an estimate of the bias, standard error and MSE of the median of the observations as estimator of θ.

b Compute a confidence interval around your estimate of the MSE that shows the size of the simulation error.

c Suppose that we reject the null hypothesis $H_0: \theta = 0$ if the median of the observations is bigger than 2. Give an accurate approximation of the size (which is very small) of this test by using (nontrivial) importance sampling. (N.B. the relevant expectation is presently a 10-dimensional integral, as there are 10 observations. This does not effect the idea of importance sampling in any way, but note that you must reweight the joint density of $(X_1, \ldots, X_{10})$, a function of 10 variables.)

Problem 4. Consider the following (overly simplified) model for the occurrence of genomic copy number variations. A genome is a string of N states each having value n ("normal") or g ("gain") modelled by the values of a Markov chain $Y_1, \ldots, Y_N$ on the state space $\{n, g\}$. The transition matrix of this Markov chain is

$$\Pi = \begin{pmatrix} p & 1-p \\ 1-q & q \end{pmatrix}.$$

We observe a noisy version $X_1, \ldots, X_N$ of this Markov chain given by the conditional (or "output") distribution

$$X_i | \, Y_i = n \sim N(\mu_n, \sigma_n^2), \qquad X_i | \, Y_i = g \sim N(m_g, \sigma_g^2).$$

Given $Y_1, \ldots, Y_N$ the outputs $X_1, \ldots, X_N$ are conditionally independent, and the conditional distirbution of X_i given $Y_1, \ldots, Y_N$ depends only on Y_i (and is as given in the display).

The observed data $X_1, \ldots, X_N$ is given in the file `cgh.txt`.

a Estimate p, q, μ_n, μ_g, and σ^2.

b Find the most likely state sequence.

c Plot the data and the most likely state sequence in a single informative figure.