Norges teknisk–
naturvitenskapelige universitet
Mathematical Sciences

Corresponding teacher:
Jo Eidsvik 90127472

EXAM IN TMA4300 COMPUTATIONAL STATISTICS

Monday 21 May, 2012

Time: 09:00–13:00

Permitted assisting material:
Yellow A5-sheet with own handwritten notes.
Tabeller og formler i statistikk (Tapir Forlag).
K. Rottmann: *Matematisk formelsamling.*
Calculator without memory function.

ENGLISH

Results of exam: 11 June, 2012.

Problem 1 Simulation

a) The Weibull distribution has cumulative function

$$F(x) = 1 - \exp(-(x/\lambda)^k), x > 0, k > 0, \lambda > 0.$$

Describe how to sample from the Weibull distribution by inversion (probability integral transform).

The exponential distribution is a special case with $k = 1$, while the Rayleigh distribution has $k = 2$. Assume that we have used inversion to sample an exponential variable X with parameter λ. Use the inversion formula to show that $Y = \sqrt{\lambda X}$ is Rayleigh distributed.

The Rayleigh distribution is used to model the time until a component fails (life time analysis). We are interested in the total time on test. Assume that 10 independent components are tested, all with Rayleigh distributed life times $X_1, \ldots, X_{10}$. The total time on test is $T = X_1 + \ldots + X_{10}$. The distribution of T can be studied by simulation of the individual life times.

b) Describe the Monte Carlo method to estimate the median in the distribution of T, that is m defined by $\int_0^m f(t)dt = 0.5$, where $f(t)$ is the density of T.

Assume the testing is done by starting one component immediately after the previous component fails. We have limited waiting capacity, and want to estimate the tail probability $p = P(T > \tau)$, where τ is a known constant related to the capacity. Often, p is small.

c) Describe the Monte Carlo method for computing an estimator $\hat{p}$ for $p = P(T > \tau)$.

What is the variance of $\hat{p}$?

The coefficient of variation (CV) is defined by the standard deviation divided by the mean. Find the CV for $\hat{p}$, and sketch how it varies with p and the number of Monte Carlo simulations.

d) We will next use importance sampling to estimate p. Assume the ten independent life times are Rayleigh distributed with parameter λ. We choose the exponential distribution with parameter λ to simulate individual, independent, life times $X_1, \ldots, X_{10}$.

Find the importance sampling estimator of p using this approach.

Problem 2 Markov chain Monte Carlo

Let y_i be the response variable at time $i = 1, \ldots, n$. Consider the linear regression model $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where x_i are known covariates at time i, $\beta = (\beta_0, \beta_1)^t$ unknown regression parameters, and ϵ_i Gaussian noise with variance $\sigma^2 = 1/q$, $q > 0$. I.e. the precision of ϵ_i is q. Because of external effects over time, the model allows the noise terms to be correlated: $\mathrm{Corr}(\epsilon_i, \epsilon_{i+h}) = \phi^h$, der $0 < \phi < 1$.

Let $f(\epsilon) = N(\epsilon; 0, \frac{1}{q}R)$ be the Gaussian density for the noise terms $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^t$. The correlation entails matrix elements (i, j) in R like $R_{i,j} = \phi^{|i-j|}$. In compact form, with data $y = (y_1, \ldots, y_n)^t$, we have

$$y = X\beta + \epsilon,$$

where X is a $n \times 2$ matrix with 1s in the first column and x_is in the second. The likelihood is :

$$f(y|\beta, q, \phi) = N(y; X\beta, \Sigma) = \frac{1}{(2\pi)^{n/2}} \frac{1}{|\Sigma|^{1/2}} \exp(-\frac{1}{2}(y - X\beta)^t \Sigma^{-1} (y - X\beta)),$$

where $\Sigma = \frac{1}{q}R$ and $\Sigma^{-1} = qR^{-1}$.

We do a statistical analysis of the model parameters, given the data. We introduce independent prior densities for the model parameters as follows: : $f(\beta) = N(\beta; 0, S)$, $\beta \in R^2$, with known S, $f(q) \propto q^{a-1} \exp(-bq)$, $q > 0$, with known a og b, and $f(\phi) = 1$, $0 < \phi < 1$.

a) Find the full conditional distribution of β.

Find the full conditional distribution of q.

b) Assume ϕ is fixed. Describe in detail how you can use a) above to construct a Gibbs sampling algorithm to simulate from the $f(\beta, q|y, \phi)$ distribution.

c) The algorithm is extended by a Metropolis-Hastings step for ϕ, given all other variables. Set $\phi(b)$ as current state of ϕ in the Markov chain, and introduce proposal distribution $g(\phi|\phi(b)) \propto \phi^{c-1}(1-\phi)^{d-1}$, for ϕ, $0 < \phi < 1$. Here $c = r\phi(b) > 0$, $d = r(1-\phi(b)) > 0$, so the mean $E(\phi|\phi(b)) = c/(c+d) = \phi(b)$. Further, r is a fixed parameter, set by us.

Find the acceptance probability of the Metropolis-Hastings step.

Describe briefly what happens with the proposal, acceptance probability and the resulting Markov chain when r is very small or large.

Problem 3 Classification and bootstrap

A Gaussian mixture distribution for vector x has density

$$f(x) = \sum_{l=1}^{K} \pi_l N(x; \mu_l, \Sigma_l), \quad \sum_{l=1}^{K} \pi_l = 1,$$

where $N(x; \mu_l, \Sigma_l)$ is the density of a Gaussian with mean μ_l and covariance matrix Σ_l. We assume $\Sigma_l = \Sigma$ for all classes $l = 1, \ldots, K$.

Two responses are measured in a lab experiment of rock samples, i.e. the response vector is $x = (x_1, x_2)^t$. We want to classify the response in one of $K = 2$ classes: resource and waste rock.

a) Describe the use of Linear Discriminant Analysis (LDA) to classify the response in two classes.

Classify data $x = (0.6, 0.2)^t$ when $\mu_1 = (0, 0)^t$, $\mu_2 = (1, 0)^t$, $\pi_1 = 0.5$ and

$$\Sigma = \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}$$

What if the 0.9 elements in the covariance matrix change to 0?

b) In practice the weight π_1, mean values and covariance matrices are estimated from data. Assume we have N bivariate lab measurements $x^1, \ldots, x^N$. Take for granted that we have a routine for computing the maximum likelihood estimator (MLE). The N data values give MLEs as in a).

Describe how to use parametric bootstrap to assess the uncertainty in the MLEs here.