Computational Statistics
6. exercise session, 26.3.2010

We will apply several techniques to the following **genetic linkage model**, which goes back to R. A. Fisher in the 1920's. 197 animals are distributed into four categories with the following frequencies

Category	1	2	3	4
Frequency	125	18	20	34

Conditionally on $\Theta = \theta$, the four categories have probabilities given by the vector

$$\left(\frac{1}{2} + \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\right),$$

where $0 < \theta < 1$. We take the uniform distribution on $(0, 1)$ as our prior. The posterior density is then proportional to the multinomial likelihood,

$$p(\theta \mid y) \propto \left(\frac{1}{2} + \frac{\theta}{4}\right)^{y_1} \left(\frac{1}{4}(1-\theta)\right)^{y_2} \left(\frac{1}{4}(1-\theta)\right)^{y_3} \left(\frac{\theta}{4}\right)^{y_4},$$

where y_i is the observed frequency of category i, and $y = (y_1, y_2, y_3, y_4)$.

We work with the following unnormalized target density

$$q(\theta) = \theta^{y_4} (1-\theta)^{y_2+y_3} (2+\theta)^{y_1} = \theta^{34} (1-\theta)^{38} (2+\theta)^{125}, \qquad 0 < \theta < 1,$$

which is proportional to the posterior density and also proportional to the likelihood. The posterior mode is

$$\hat{\theta} = \frac{15 + \sqrt{53809}}{394} \approx 0.6268214980.$$

The normalizing constant can be found exactly using computer algebra. It is

$$\int_0^1 q(\theta)\, \mathrm{d}\theta = 0.2357695164567474 \cdot 10^{29}$$

with sixteen significant digits. The derivatives of $L(\theta) = \log q(\theta)$ are given by

$$L'(\theta) = \frac{-197\,\theta^2 + 15\,\theta + 68}{\theta(1-\theta)(2+\theta)}$$

and

$$L''(\theta) = \frac{-197\,\theta^4 + 30\,\theta^3 - 175\,\theta^2 + 136\,\theta - 136}{\theta^2(1-\theta)^2(2+\theta)^2}.$$

1. Approximate the normalizing constant of the posterior in the genetic linkage example by the grid method. First define a function, say `upost`, to evaluate the unnormalized posterior density. Then you can define the grid and evaluate the function on the grid as follows in R.

```
h <- 1/N
tgrid <- seq(h/2, 1 - h/2, length = N)
upost(tgrid)
```

Compare the grid method approximation to the exact value of the normalizing constant, when N is 10, 20, 40 and 80. (In R, the numbers are by default printed only with seven significant digits. In order to see variable `res`, with ten significant digits, call `print(res, digits = 10)`).

R has the function `integrate` for one-dimensional numerical integration. Try, what you get with `integrate(upost, 0, 1)`

2. Now we apply Laplace's method to approximate the normalizing constant $c = \int q(\theta)\,\mathrm{d}\theta$ of an unnormalized univariate posterior $q(\theta) = q(\theta \mid y)$. We first form the second degree Taylor approximation for $\log q(\theta)$ centered at its mode $\hat{\theta}$,

$$\log q(\theta) \approx \log q(\hat{\theta}) - \frac{1}{2} A(\theta - \hat{\theta})^2.$$

Exponentiating, we form an approximation to $q(\theta)$, which we then integrate over the whole real line with respect to θ in order to find the approximation $\tilde{c}$ to c.

a) Give a formula for $\tilde{c}$. Evaluate $\tilde{c}$ in the genetic linkage example.

b) What is the connection of the following calculation and of the normal approximation (6.4)?

$$\frac{1}{\tilde{c}} \exp\left(\log q(\hat{\theta}) - \frac{1}{2} A(\theta - \hat{\theta})^2\right)$$

3. Evaluate the two Laplace approximations, eq. (6.11) and eq. (6.12), for the posterior expectation of Θ in the genetic linkage example.

4. Once again, consider $E[\Theta \mid y]$ in the genetic linkage example. Calculate it

a) by numerical integration,

b) by Monte Carlo integration, where you use one of the following methods: the grid method, accept–reject (use the prior as the proposal), or importance sampling (with instrumental density equal to prior).

5. When the statistical model includes a discrete parameter, then we often need to handle sums or ratios of the form

$$s = \sum_{j=1}^{k} q_j, \qquad p_i = \frac{q_i}{\sum_{j=1}^{K} q_j}, \quad i = 1, \ldots, K$$

where $q_j > 0$ are positive numbers. Sometimes the numbers q_j are so small, that they cannot be represented as floating point numbers, whereas their logarithms can. So,

$$z_j = \log q_j \quad \Leftrightarrow \quad q_j = \exp(z_j).$$

Find a way to calculate $\log(s)$ and the numbers $p_1, \ldots, p_K$ under such a situation. Hint: take the largest of the q_j:s as a common factor.

In particular, calculate $\log(s)$ and p_1, p_2, p_3, when

$$(z_1, z_2, z_3) = (-1000, -1001, -1002).$$

(In this case the floating point representation of each $\exp(z_i)$ is 0, at least when you use double precision floating point numbers, as R does.)