Norwegian University of Science and Technology
Department of Mathematical Sciences

English

Contact during exam:
Inge Myrseth 73 59 04 84

EXAM IN TMA4300 MODERN STATISTICAL METHODS

Wednesday May 27th 2009
Time: 09:00–13:00

Aids: Calculator HP30S.
Statistiske tabeller og formler, Tapir forlag.
K. Rottman: Matematisk formelsamling.
One yellow paper (A4 with stamp) with own formulas and notes.

Grading: June 17th 2009.

Problem 1

Consider a random sample $X_1, \ldots, X_n$ from the normal distribution with (unknown) mean μ and (known) variance $1/\tau^2$, i.e. with density function

$$f(x;\mu) = \frac{\tau}{\sqrt{2\pi}} \exp\left\{-\frac{\tau^2}{2}(x-\mu)^2\right\}.$$

a) Write $f(x;\mu)$ on the form of a one-parameter exponential family,

$$f(x;\mu) = a(x)e^{\phi(\mu)t(x)+b(\mu)}$$

i.e. identify the functions $a(x)$, $\phi(\mu)$, $t(x)$ and $b(\mu)$.

Use this to write down a formula for the conjugate prior distribution for μ. Show that the conjugate prior for μ is a normal distribution.

b) Using the prior $\mu \sim \mathrm{N}(\nu, 1/r^2)$, show that the posterior distribution $f(\mu|x_1, \ldots, x_n)$ is also a normal distribution. In particular, express the posterior mean, $\mathrm{E}[\mu|x_1, \ldots, x_n]$, and the posterior variance, $\mathrm{Var}[\mu|x_1, \ldots, x_n]$, in terms of ν, r, τ, n and $x_1, \ldots, x_n$.

Problem 2

In this problem we will study mortality rates in eleven hospitals when performing a particular type of surgery. The observed data is given in Table 1 below. We model the number of deaths for hospital H_i, x_i, with a binomial distributions with parameters n_i and p_i, where n_i is the number of operations at that hospital and p_i is an unknown probability of death for hospital H_i. Moreover we assume $x_1, \ldots, x_{11}$ to be conditionally independent for given parameter values $p_1, \ldots, p_{11}$.

For each $i = 1, \ldots, 11$, define

$$b_i = \ln\left(\frac{p_i}{1-p_i}\right) \quad \Leftrightarrow \quad p_i = \frac{e^{b_i}}{1+e^{b_i}},$$

so that $b_i \in \mathbb{R} \Leftrightarrow p_i \in (0,1)$. Furthermore, apriori we assume $b_1, \ldots, b_{11}$ to be independent given two (unknown) hyper-parameters μ and τ^2, and

$$b_i \sim \mathrm{N}(\mu, 1/\tau^2) \text{ for } i = 1, \ldots, 11.$$

Finally, we assume μ and τ^2 to be apriori independent, $\mu \sim \mathrm{N}(\nu, 1/r^2)$ and that τ^2 has a gamma distribution with parameters α and β, i.e.

$$f(\tau^2) = \frac{1}{\beta^\alpha \Gamma(\alpha)} \left(\tau^2\right)^{\alpha-1} \exp\left\{-\frac{\tau^2}{\beta}\right\} \text{ for } \tau^2 > 0.$$

We assume ν, r, α and β to have known values.

Our task in this problem is to define a single site Metropolis–Hastings algorithm to sample from the posterior distribution $f(b_1, \ldots, b_{11}, \mu, \tau^2|x_1, \ldots, x_n)$ and then use the generated samples to estimate properties of that distribution.

Hospital	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8	H_9	H_{10}	H_{11}
No. of operations	148	119	810	211	196	148	215	207	97	256	360
No. of deaths	18	8	46	8	13	9	31	14	8	29	24

Table 1: Number of operations, n_i, and number of deaths, x_i, for each of eleven hospitals.

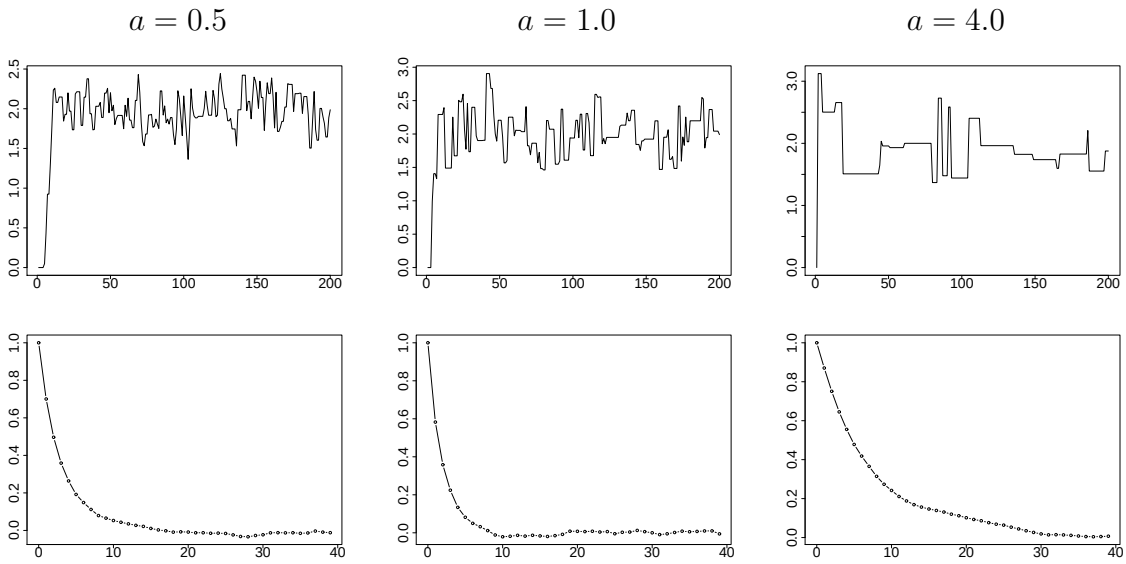

Figure 1: Trace plots (upper row) of the first 200 iterations for b_1 and corresponding estimated autocorrelation functions (lower row) for simulated b_1-values after convergence for $a = 0.5$ (left column), $a = 1.0$ (middle column) and $a = 4.0$ (right column).

a) Visualise the Bayesian model specified above as a graphical model.

Explain how the results in Problem 1 can be used to define a proposal distribution for μ. What is the corresponding acceptance probability?

Find the full conditional distribution for τ^2. In particular show that this is a gamma distribution and identify formulas for the two parameters of this gamma distribution.

A random walk proposal is adopted for each b_i. A potential new value, $\widetilde{b}_i$, is generated from a uniform distribution centered at the current value b_i, i.e.

$$\widetilde{b}_i \sim \text{Unif}[b_i - a, b_i + a],$$

where a is a tuning parameter.

b) Find an expression for the acceptance probability for $\widetilde{b}_i$. The expression should be simplified as much as possible.

Explain why it would be a waste of computation time to use the proposal distribution $\widetilde{b}_i \sim \text{Unif}[b_i - 2a, b_i + a]$.

In Figure 1 trace plots and estimated autocorrelation functions for p_1 is shown for three different values of the tuning parameter a.

c) Based on what you can see from these plots, which of the three values for a would be prefer? Give reason(s) for your answer.

Are there other plots or values you would study before choosing a good value for the tuning parameter a? If yes, explain which one and how you would use it to choose a value for a.

Assume we have run the MCMC algorithm for M iterations and denote the generated states by $\{\mu^m, \tau^m, b_1^m, \ldots, b_{11}^m\}_{m=0}^M$. In particular $\mu^0, \tau^0, b_1^0, \ldots, b_{11}^0$ is the initial state.

d) Specify how you from the simulated values will estimate the following three quantities.

1. $\mathrm{E}[p_i|x_1, \ldots, x_{11}]$.
2. $\mathrm{Prob}(p_i < p_j|x_1, \ldots, x_{11})$ for $i \neq j$.
3. $\mathrm{Prob}(p_i < \min_{j\neq i} p_j|x_1, \ldots, x_{11})$.

Discuss what quantity you would prefer to decide which hospital is the best hospital for this kind of surgery.

Problem 3

In this problem we will consider the same data set as in Problem 2, but we will now use parametric bootstrapping to analyse the situation. Again we assume that $x_1, \ldots, x_{11}$ are independent and

$$x_i \sim \mathrm{Binomial}(n_i, p_i),$$

for some unknown parameters $p_1, \ldots, p_{11}$. To estimate p_i we use

$$\widehat{p}_i = \frac{x_i}{n_i}.$$

a) Considering first hospital H_i only, define the ideal bootstrap estimator for the standard deviation of $\widehat{p}_i$ (based on parametric bootstrapping).

In this simple situation with only one hospital it is possible to find the ideal bootstrap estimator for standard deviation of $\widehat{p}_i$ analytically (i.e. without simulation). Find a simple expression for this ideal bootstrap estimator.

Hospital	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8	H_9	H_{10}	H_{11}
$\widehat{p}_i$	0.122	0.0672	0.057	0.038	0.0663	0.061	0.144	0.068	0.082	0.113	0.0667
$\widehat{r}_i$	10	6	2	1	4	3	11	7	8	9	5

Table 2: Parameter estimates $\widehat{p}_1, \ldots, \widehat{p}_{11}$ and corresponding estimated ranks, $\widehat{r}_i$, found from the data in Table 1.

Based on the data in Table 1 we get estimates $\widehat{p}_1, \ldots, \widehat{p}_{11}$ as given in Table 2. Define $r_i, i = 1, \ldots, 11$ to be the rank of hospital H_i based on (the unknown) $p_1, \ldots, p_{11}$, i.e. $r_i = 1$ for the hospital with the lowest value for p_i, $r_i = 2$ for the hospital with the second smallest value for p_i and so on. In particular $r_i = 11$ for the hospital with the highest value for p_i. We estimate r_i by the corresponding rank based on the estimated parameter values $\widehat{p}_1, \ldots, \widehat{p}_{11}$. The estimated values $\widehat{r}_1, \ldots, \widehat{r}_{11}$ are also given in Table 2.

We now want to use parametric bootstrapping to find a (percentile) confidence interval for each of $r_1, \ldots, r_{11}$.

b) Write pseudo code for finding a percentile confidence interval for each $r_1, \ldots, r_{11}$.

Clearly r_i can only take integer values, does this have any implications for how to obtain the confidence interval? Give reason(s) for your answer.