UNIVERSITETET I OSLO

Matematisk Institutt

EXAM IN:	**STK 4021/9021 – Bayesian statistics**
	Part II of two parts
WITH:	**Nils Lid Hjort**
AUXILIA:	**Calculator, plus one single sheet of paper**
	with the candidate's own personal notes
TIME FOR EXAM:	**Part I: The Project, 6–19/xii/2013;**
	Part II: 9/xii s.y., 9:00–13:00, written exam

This exam set contains four exercises and comprises four pages. The last page also provides a short Appendix containing a few facts and definitions which you may find useful.

Note: Please write **your name** on the top of the first page of what you hand in today. In the marking process we need to connect your solutions of today with your project report.

Exercise 1

WE START OFF considering a simple but prototypical Bayesian setup, involving a normal model with a normal prior. Suppose y is a single observation from the $\mathrm{N}(\theta, 1)$ distribution, with θ being given a $\mathrm{N}(\theta_0, \sigma_0^2)$ prior.

(a) What is then the marginal distribution of y, and what is the covariance between y and θ?

(b) Find also the posterior distribution of θ, expressed in terms of the observation y. Comment specifically but briefly on the two cases where σ_0 is respectively small or big.

(c) On this occasion Mr Savage uses the prior $\mathrm{N}(-1, 0.780^2)$ for θ, reflecting in particular his being rather sure that the θ in question is *negative*, as $\Pr\{\theta < 0\} = \Phi(1.282) = 0.90$ (writing as usual $\Phi(\cdot)$ for the cumulative standard normal distribution). Write up an expression for Mr Savage's belief that θ is negative after having seen the datum y. How big must this observation y be in order for Mr Savage to change his mind into then believing that θ is *positive* with probability 0.90?

Exercise 2

LORD RAYLEIGH WON THE NOBEL PRIZE for physics in 1904 (jointly with William Ramsey, for having discovered argon). A certain statistical distribution is named after him, and is used in various branches of statistics, e.g. in connection with analysis of waves, directions, and amplitudes, and with magnetic resonance images. Say now that y has the Rayleigh distribution with parameter θ, written $y \sim \mathrm{Rayl}(\theta)$, if its density is

$$f(y, \theta) = \exp(-\tfrac{1}{2}\theta y^2)\theta y \quad \text{for } y > 0,$$

where θ is a positive parameter.

(a) Assume independent data points $y_1, \ldots, y_n$ are observed from the Rayl(θ) distribution. Find the maximum likelihood estimator $\widehat{\theta}$, expressed in terms of $w_n = n^{-1} \sum_{i=1}^n y_i^2$. Also exhibit a normal approximation to the distribution of $\widehat{\theta}$.

(b) Suppose further that prior knowledge about θ may be adequately translated into the prior $\theta \sim \text{Gam}(a, b)$, for suitable values of (a, b) (see the Appendix here for the definition of the Gamma distribution). Show that the posterior distribution of θ also is of the Gamma type, say $\theta \,|\, \text{data} \sim \text{Gam}(a_n, b_n)$, and identify a_n, b_n. What can you say here regarding this exact posterior distribution in relation to the so-called 'Lazy Bayesian's approximation'?

(c) Give a formula for the Bayes estimate of θ under squared error loss. Explain briefly how you may construct a 90% credibility interval for θ.

(d) For the Gam(a, b) prior, find the predictive distribution $\bar{f}(y)$, the distribution of 'the next data point' y_{n+1}, given the observed data $y_1, \ldots, y_n$.

Exercise 3

HERE WE SHALL BRIEFLY CONSIDER a type of 'statistical pattern recognition' problem found in e.g. various types of communication theory.

(a) Suppose there are two competing models for explaining a data point (or vector) y; under model M_1 the density is $f_1(y)$, and under M_2 it is $f_2(y)$. For simplicity we take these model densities to be known, i.e. without further unknown parameters. If one in addition has prior probabilities π_1 and π_2 for these two models (summing to 1, i.e. no other models are under consideration), explain why

$$\Pr(M_1 \,|\, y) = \frac{\pi_1}{\pi_1 + \pi_2 R(y)}, \quad \Pr(M_2 \,|\, y) = \frac{\pi_2 R(y)}{\pi_1 + \pi_2 R(y)},$$

where $R(y) = f_2(y)/f_1(y)$.

(b) Assume a decision needs to be taken after having observed y, either A_1, 'the y stems from f_1' or A_2, 'the y stems from f_2', and that the loss function is involved is

$$L(\text{model}, \text{decision}) = \begin{cases} 1 & \text{if one is wrong,} \\ 0 & \text{if one is correct.} \end{cases}$$

Derive the Bayes rule for this problem, that is, explain clearly when the Bayes rule allocates y to M_1 and when to M_2.

(c) Consider the two simple densities

$$f_1(y) = \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2} y^2) \quad \text{and} \quad f_2(y) = \frac{1}{\sqrt{2}} \exp(-\sqrt{2}|y|).$$

The first is of course the standard normal and the second is a so-called double exponential, which I have scaled here to have standard deviation 1. The two densities are hence symmetric with identical means and identical standard deviations and not easy

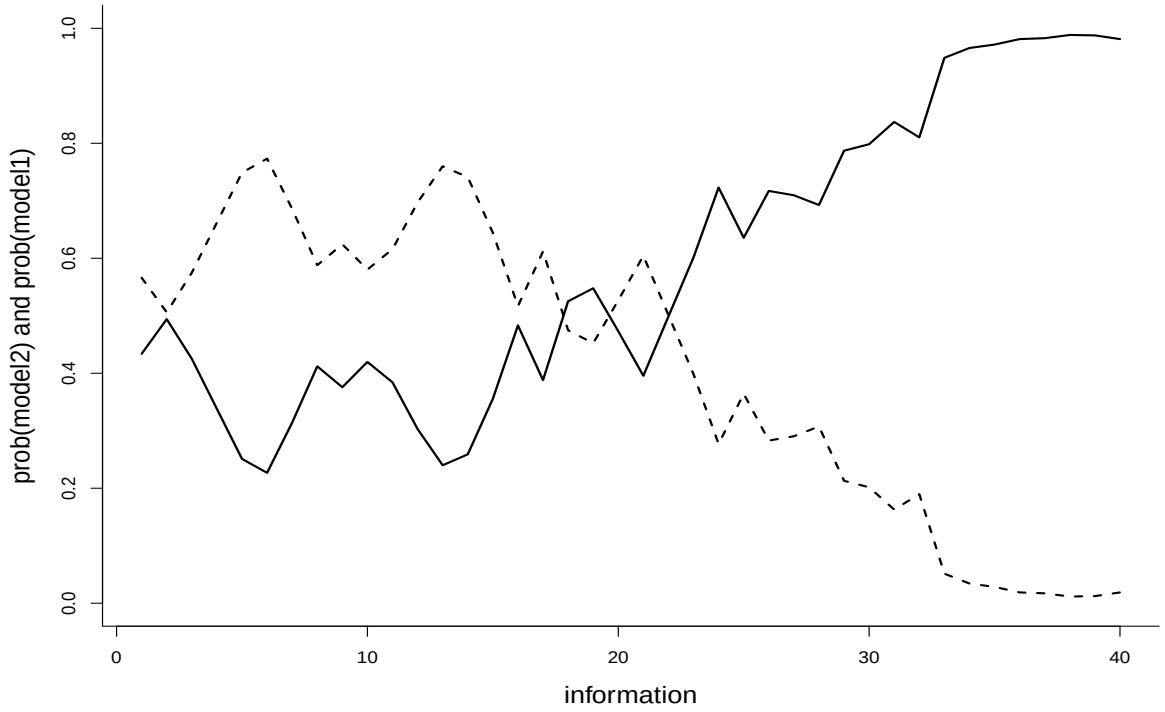
to disentangle from each other. With equal prior model probabilities $\pi_1 = \pi_2 = \frac{1}{2}$, find out precisely when an observed y is classified as standard normal and when it is classified as double exponential.

(d) The problem becomes statistically speaking easier with more information about the underlying source. Assume in fact that you are observing a stream of independent data $y_1, y_2, y_3, \ldots$, where these are either all stemming from f_1 or all from f_2, with equal prior probabilities $\frac{1}{2}, \frac{1}{2}$. Give formulae for

$$\Pr(M_1 \,|\, y_1, \ldots, y_j) \quad \text{and} \quad \Pr(M_2 \,|\, y_1, \ldots, y_j)$$

for $j = 1, 2, 3, \ldots$. – This is illustrated in the figure below, based on having observed so far forty data points:

```
 1.982  0.257  0.867 -1.077 -1.294  0.648  0.092 -0.108 -0.672  0.308
 2.162  1.049 -0.950  2.443 -0.084  0.032 -1.134  0.011 -0.396 -1.920
 1.887 -0.113  0.115 -0.018  1.192  0.148  0.530  0.582 -0.056 -0.418
 0.241  0.718 -3.360 -0.110 -0.297 -0.113 -0.398  0.128  0.559 -1.356
```



The figure displays $\Pr(M_2 \,|\, y_1, \ldots, y_j)$ *(full curve) and* $\Pr(M_1 \,|\, y_1, \ldots, y_j)$ *(dotted line) as information about the source accumulates, for* $j = 1, 2, 3, \ldots$.

(e) Attempt to prove that the machinery developed above is guaranteed to work correctly with enough data, in the sense that if the data come from f_1, then $\Pr(M_1 \,|\, \text{data}) \to 1$, and correspondingly that if the data come from f_2, then $\Pr(M_2 \,|\, \text{data}) \to 1$. The convergence here is for the sample size increasing towards infinity.

Exercise 4

CONSIDER THE ORDERED measurements

0.255 0.818 0.859 2.504 2.549 2.793 3.039 3.603 3.805 4.294

These are taken to be an ordered i.i.d. sample from a uniform distribution on $[0, \theta]$, with θ an unknown parameter.

(a) Write down the likelihood function for these data, under the assumed uniform model. What is the maximum likelihood estimate?

(b) With the prior $1/\theta$ for the unknown parameter, give an explicit formula for the posterior distribution, and compute the Bayes estimate under absolute loss (i.e. $L(\theta, \widehat{\theta}) = |\widehat{\theta} - \theta|$).

Appendix: Just a few useful facts

The multinormal distribution: An important property of the multinormal is that a subset of components, conditional on another subset of components, remains multinormal. In fact, if

$$X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix} \sim \mathrm{N}_{k_1+k_2}(\begin{pmatrix} \xi^{(1)} \\ \xi^{(2)} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}),$$

then

$$X^{(1)} \mid \{X^{(2)} = x^{(2)}\} \sim \mathrm{N}_{k_1}(\xi^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}(x^{(2)} - \xi^{(2)}), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}).$$

The gamma distribution: We say that X has a Gamma distribution with parameters (a, b), which we write as $X \sim \mathrm{Gam}(a, b)$, if its density takes the form

$$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx) \quad \text{for } x > 0.$$

Its mean and variance are equal to respectively a/b and a/b^2.