UNIVERSITETET I OSLO

Matematisk Institutt

EXAM IN:	**STK 4021/9021 – Applied Bayesian Analysis**
WITH:	**Nils Lid Hjort**
AUXILIA:	**Calculator, plus one single sheet of paper with the candidate's own personal notes**
TIME FOR EXAM:	**Tuesday 30/xi/2021, 9:00 – 13:00**

This exam set contains four exercises and comprises five pages, with the fifth and last page being a simple appendix.

Exercise 1: estimating many normal means

HERE WE SHALL CONSIDER a simple normal prototype setup, first the basic one where we observe a single y which given θ comes from a $\mathrm{N}(\theta, 1)$ distribution.

(a) The classic estimator for the mean parameter is of course $\widehat{\theta}_c = y$ itself. Find its risk function, under squared error loss, i.e. $r_c(\theta) = \mathrm{E}_\theta\,(y-\theta)^2$.

(b) Then assume the prior distribution $\theta \sim \mathrm{N}(0, \sigma^2)$, with σ given. Show that the joint distribution, for the parameter and the data point, is the binormal

$$\begin{pmatrix} \theta \\ y \end{pmatrix} \sim \mathrm{N}_2\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \sigma^2 \\ \sigma^2 & \sigma^2+1 \end{pmatrix}\Big).$$

– I do not wish you to spend too many exam minutes on this point, so it is sufficient that you explicitly find the means, the variances, and the covariance (then joint binormality follows via a few extra arguments).

(c) You are not required to show this here and now, but it is an easy task to deduce from the joint distribution above that

$$\theta \,|\, y \sim \mathrm{N}(\rho y, \rho), \quad \text{with } \rho = \frac{\sigma^2}{\sigma^2+1}.$$

But show that the posterior mean estimator (i.e. the Bayes estimator under quadratic loss), is $\widehat{\theta}_B = \rho y$, and find its risk function $r_B(\theta)$. Find also the interval of θ inside which the Bayes risk function is smaller than the classic risk function.

(d) Assume now that there are independent normal observations $y_1, \ldots, y_{100}$ of the above type, with different mean parameters and variances known to be 1, i.e. $y_i \,|\, \theta_i \sim \mathrm{N}(\theta_i, 1)$. If one uses the same prior for all these mean parameters, with $\theta_1, \ldots, \theta_{100}$ being independent from $\mathrm{N}(0, \sigma^2)$, then the recipe above leads to

$$\widehat{\theta}_i = \mathrm{E}\,(\theta_i \,|\, y_i) = \rho y_i = \frac{\sigma^2}{\sigma^2+1} y_i \quad \text{for } i = 1, \ldots, 100.$$

This requires σ to be known, however, which is most often not realistic. Construct therefore a new estimator, of the type

$$\theta_i^* = \widehat{\rho} y_i \quad \text{for } i = 1, \ldots, 100,$$

where you propose a way of estimating ρ from the hundred observations.

Exercise 2: count me in

CERTAIN COUNT VARIABLES y follow the distribution given by

$$f(y,\theta) = \Pr(Y = y \,|\, \theta) = (y-1)(1-\theta)^{y-2}\theta^2 \quad \text{for } y = 2,3,4,\ldots,$$

where θ is a parameter in $(0,1)$. You do need to show this here, but with a bit of patience one can show that

$$\mathrm{E}\,(y \,|\, \theta) = \frac{2}{\theta}, \quad \mathrm{Var}\,(y \,|\, \theta) = \frac{2(1-\theta)}{\theta^2}.$$

(a) Assume first, in the spirit of Bayes's 1763 essay, that not knowing θ is formalised as meaning that it stems from a uniform distribution on $(0,1)$. Find the posterior distribution $\theta \,|\, y$.

(b) Still taking θ to have a uniform prior, find also the implied marginal distribution for y, i.e. a formula for $\Pr(Y = y)$ for $y = 2,3,4,\ldots$. Find the mean $\mathrm{E}\,Y$ for this distribution.

(c) Suppose now that independent count data $y_1,\ldots,y_n$ are observed from the model $f(y,\theta)$, with the same θ. Write down the likelihood function, and show that the maximum likelihood estimator is $\widehat{\theta}_{\mathrm{ml}} = 2/\bar{y}$, in terms of the data mean $\bar{y} = (1/n)\sum_{i=1}^n y_i$.

(d) With any given smooth prior for θ, explain what the so-called Lazy Bayes posterior distribution is here, based on a normal approximation.

(e) Regardless of such a normal approximation to a given posterior distribution, assume now that the prior for θ is a Beta(a,b) (see the Appendix). Find the exact posterior distribution for θ, and show that

$$\mathrm{E}\,(\theta \,|\, \text{data}) = \frac{a+2n}{a+b+n\bar{y}}.$$

(f) Suppose that ten data points

3, 7, 4, 6, 9, 23, 10, 13, 8, 7

have been observed, and that the uniform prior is used for θ. What is then the distribution of the eleventh data point, not yet observed? The task is to find a formula for $\Pr(Y_{11} = y \,|\, y_1,\ldots,y_{10})$, for $y = 2,3,4,\ldots$.

Exercise 3: are you Type 1, or are you Type 2?

A RATHER SIMPLIFIED SETUP for statistical pattern recognition (classification analysis, discriminant analysis) is as follows. There are only two classes, 1 and 2, and a single observation y, coming from density $f_1(y)$ if from class 1, and density $f_2(y)$ if from class 2. The task is to classify y as coming from class 1 or class 2. The y can be multidimensional, or a sequence, or an image (is it a cat?, is it a dog?), without disturbing the basic setup here.

Suppose here that the two classes are equally likely a priori (if you need notation, take $\pi_1 = \frac{1}{2}$, $\pi_2 = \frac{1}{2}$), and that $f_1(y)$ and $f_2(y)$ are completely specified, so there is only one unknown parameter in the game, namely the class label $c \in \{1, 2\}$. Suppose further that the statistician is to reach a decision $\widehat{c}$, for the observed y, with three possibilities, $\{1, 2, D\}$. Here $\widehat{c} = 1$ means that one claims y is of type 1, similarly for $\widehat{c} = 2$, whereas decision D means one is in doubt, perhaps to be followed by further inspection, another measurement for the same object, or the like. The loss function is here taken to be

$$L(c, \widehat{c}) = \begin{cases} 0 & \text{if correct,} \\ 1 & \text{if wrong,} \\ 0.10 & \text{if } D. \end{cases}$$

(a) Show first that

$$\Pr(c = 1 \,|\, y) = \frac{f_1(y)}{f_1(y) + f_2(y)} = \frac{1}{1 + R(y)},$$
$$\Pr(c = 2 \,|\, y) = \frac{f_2(y)}{f_1(y) + f_2(y)} = \frac{R(y)}{1 + R(y)},$$

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

(b) Show next that

$$\mathrm{E}\,(L(c, 1) \,|\, y) = 1 - \Pr(c = 1 \,|\, y),$$
$$\mathrm{E}\,(L(c, 2) \,|\, y) = 1 - \Pr(c = 2 \,|\, y).$$

(c) Then show that the Bayes solution to the classification problem is

$$\text{claim 1 if } \Pr(c = 1 \,|\, y) \geq 0.90,$$
$$\text{claim 2 if } \Pr(c = 2 \,|\, y) \geq 0.90,$$
$$\text{use } D \text{ if both } \Pr(c = 1 \,|\, y) < 0.90 \text{ and } \Pr(c = 2 \,|\, y) < 0.90.$$

In which precise sense is this the optimal procedure?

(d) As an illustration of the general setup here, suppose that y is one-dimensional, with $f_1(y)$ being $\mathrm{N}(-1, 1)$ and $f_2(y)$ being $\mathrm{N}(1, 1)$. Identify the three regions of y, where one claims 1, claims 2, or uses the D option.

Exercise 4: ratio of Cauchy parameters

Consider the two simple datasets

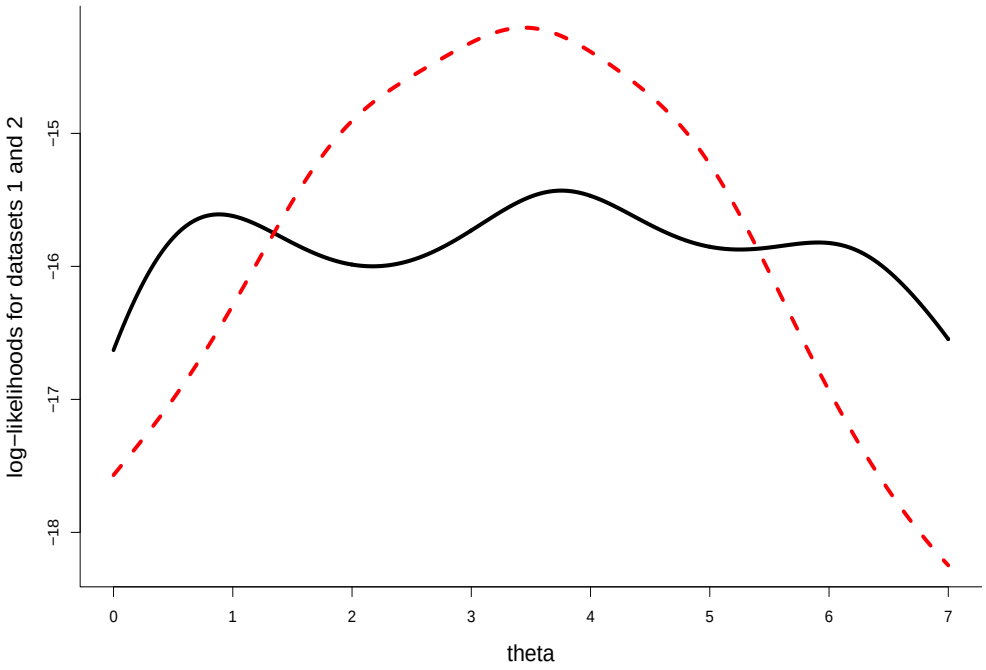
data 1: $0.158, 0.716, 3.722, 6.240, 7.632$

data 2: $-0.636, 1.846, 3.461, 4.954, 8.182$

We assume here that these are ordered independent samples from two different Cauchy models, with parameters θ_1 and θ_2, i.e. with densities

$$f(y, \theta_1) = \frac{1}{\pi} \frac{1}{1 + (y - \theta_1)^2} \quad \text{and} \quad f(y, \theta_2) = \frac{1}{\pi} \frac{1}{1 + (y - \theta_2)^2}$$

on the real line. I have plotted the two consequent log-likelihood functions, say $\ell_1(\theta_1)$ (smooth black) and $\ell_2(\theta_2)$ (dashed red), in the figure below, from zero and upwards.



The two log-likelihood functions, $\ell_1(\theta_1)$ (black, full) and $\ell_2(\theta_2)$ (red, dashed), for the two small Cauchy datasets.

(a) Here we're Bayesians, and use a flat prior on $[0, 100]$ for both θ_1 and θ_2. Write down expressions for the posterior distributions of θ_1 and θ_2, say $\pi_1(\theta_1 \,|\, \text{data}_1)$ and $\pi_2(\theta_2 \,|\, \text{data}_2)$, and explain how you would compute these (when given time to actually do so, after the exam).

(b) Then explain how you would go about sampling a high number of θ_1 and θ_2 from their respective posterior distributions.

(c) Finally explain how you can make Bayesian inference for the ratio parameter $\rho = \theta_2/\theta_1$.

Appendix: A few facts for the Beta distribution

We say that X is a Beta distribution with parameters (a, b), these being positive, provided its density is

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} \quad \text{for } x \in (0,1).$$

Here $\Gamma(\cdot)$ is the gamma function. In particular,

$$\int_0^1 x^{a-1}(1-x)^{b-1}\,\mathrm{d}x = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

and it also follows from this that

$$\int_0^1 x^m(1-x)^n\,\mathrm{d}x = \frac{m!\,n!}{(m+n+1)!}$$

when m and n are integers. Furthermore,

$$\mathrm{E}\,X = x_0 = \frac{a}{a+b} \quad \text{and} \quad \mathrm{Var}\,X = \frac{x_0(1-x_0)}{a+b+1}.$$

One may also find use for the inverse mean, which is

$$\mathrm{E}\,\frac{1}{X} = \frac{a+b-1}{a-1} \quad \text{for } a > 1.$$