UPPSALA UNIVERSITY
Department of Mathematics
Silvelyn Zwanzig

Bayesian Statistics , 10hp, 1MS900
April 12, 2022

Permitted aids: pocket calculator, one hand-written sheet of formulae (2 pages)

Time: 5 hours. For a pass (mark 3) the requirement is at least 18 points. For the mark 4, 25-31 points are necessary. For an excellent test (mark 5) the requirement is at least 32 points. Every problem is worth 5 points.

OBS: *Please explain and interpret your approach carefully. Don't try to write more than really needed, but what you write must be clear and well argued.*

1. Consider the following statistical (toy :-)) model for costumer satisfaction. In a query the costumer can rate between satisfied (+), disappointed (-), ok(+/-), no answer (0). The parameter θ is the level of satisfaction $0, 1, 2$, where 2 means the costumer was satisfied.

   The probability $P_\theta(x)$ is given in the following table:

x	$+$	$+/-$	$-$	0
$\theta = 2$	0.6	0.1	0	0.3
$\theta = 1$	0.1	0.2	0.1	0.6
$\theta = 0$	0	0.1	0.8	0.01

   From earlier studies it is known that the probability of high satisfaction ($\theta = 2$) is 0.2 and the probability of no satisfaction ($\theta = 0$) is 0.3. This model was applied to costumers using the camping place "Nice Lake" in Dalarna. Assume we have the result from n tourists, who answered independently. $\sum_{i=1}^{4} n_i = n$

answer	$+$	$+/-$	$-$	0
number	n_1	n_2	n_3	n_4

   (a) Calculate the posterior distribution.

   (b) Calculate the MAP estimator.

   (c) Compare this estimator with the maximum likelihood estimator.

2. Suppose another more simplified query for analyzing costumer satisfaction. The parameter has only two levels $\vartheta \in \{0, 1\}$ where 0 means that the costumer is satisfied. In the query one costumer can choose between (bad, acceptable, good) coded as $-1, 0, 1$. The goal is to estimate the parameter ϑ. The $0-1$ loss is assumed. The probability $P_\vartheta(x)$ is given in the following table:

x	-1	0	1
$\vartheta = 0$	0.1	0.2	0
$\vartheta = 1$	0.8	0.2	0

   .

Assume that the prior probability of $\vartheta = 1$ is p.

(a) Determine the posterior distribution.

(b) Calculate the posterior loss.

(c) Determine the Bayes estimator of ϑ.

(d) Calculate the Bayes risk.

(e) Determine the least favorable prior.

3. Consider an i.i.d sample $X_1, \ldots, X_n$ from a Poisson-distribution with parameter θ.

$$P_\theta(k) = \frac{k!}{\theta^k} \exp(-\theta)$$

(a) Belongs the sample distribution to an exponential family? Determine the sufficient statistic and the natural parameter.

(b) Derive the Fisher information.

(c) Apply the result (Prop 3.3.13) on exponential families for deriving a conjugate family.

(d) Which family of distributions is this conjugate family?

(e) Give the conjugate posterior distribution $p(\theta \mid X_1, \ldots, X_n)$.

(f) Derive Jeffreys prior.

(g) Belongs Jeffreys prior to the conjugate family?

4. The observations belong to independent random variables $\mathbf{X} = (X_{i,j})$, where

$$X_{ij} \sim N(\theta_i, 1),\ i = 1, \ldots, n,\ j = 1, \ldots, k.$$

The parameters $\theta_1, \ldots, \theta_n$ are independent and normal distributed with

$$\theta_i \sim N(\mu, 1), \quad \mu \sim N(0, 1).$$

(a) Determine the prior for $\theta = (\theta_1, \ldots, \theta_n)$.

(b) Calculate the posterior distribution θ given $\mathbf{X}$.

(c) Calculate the posterior distribution of $\bar{\theta} = \frac{1}{n} \sum_{i=1}^{n} \theta_i$ given $\mathbf{X}$.

(d) Propose a Bayes test for

$$H_0 : -0.2 \leq \bar{\theta} \leq 0.2 \quad H_1 : else.$$

5. In a medical study 1000 patients were randomly chosen and their blood are tested for multiple drug resistance, 15 persons got a positive test result. From earlier studies it is known that θ, the probability of a positive test, is around 2%. Let X be the number of positive tested patients. Two different Bayes models are proposed.

Model 1

$$X \sim Bin(1000, p), \quad p \sim Beta(1, 20)$$

Model 2

$$X \sim Pois(\lambda), \quad \lambda \sim Gamma(20, 1)$$

(a) Compare the maximum likelihood estimators of θ in both models.

(b) Give the posterior distribution in Model (1) and the Bayes estimator of p (respectively to the quadratic loss).

(c) Give the posterior distribution in Model (2) and the Bayes estimator of λ (respectively to the quadratic loss).

(d) Compare the Bayes estimators of θ.

(e) Give the formulary of the Bayes factor B_{12}.

(f) Calculating B_{12} with help of R gives the value $B_{12} = 0.272$. Compare the models.

6. In Problem 5 the Bayes factor is calculated. Here are R codes for 4 different methods for calculating one of the needed integrals.
Method 1

```
x1<-runif(10000); sum(dbeta(x1,1,20)*dbinom(15,1000,x1))/10000
```

Method 2

```
x2<-rbeta(10000,1,20); sum(dbinom(15,1000,x2))/10000
```

Method 3

```
f1<-function(x){dbinom(15,1000,x)*dbeta(x,1,20)}; integrate(f1,0,1)
```

Method 4

```
M4<-function(a,seed,N){rand<-rep(NA,N);rand[1]<-seed;
for(i in 2:N) {rand[i]<-runif(1);
r1<-(1-rand[i])/(1-seed);
r<-min(1,r1^19);
if (runif(1)<r){seed<-rand[i]}else{rand[i]<-seed}};
 return(rand)}
XX<-M4(0.1,0.5,10100)
mean(dbinom(15,1000,XX[100:10100]))
```

Following results are obtained

$Method$	1	2	3	4
$result$	0.01228	0.01397	0.01476	0.01547

(a) Which integral is calculated?

(b) Give the name of each methods.

(c) Give the steps of the algorithm of Method 2.

(d) Give the steps of the algorithm of Method 4.

(e) Why are the results so different? Which result do you prefer most? Why?

(f) Compare the accuracy of Method 1 and Method 2.

7. Consider a linear model with fixed design and independent standard normal distributed errors,

$$y_i = \alpha + \beta x_i + \epsilon_i, \quad x_1 = x_2 = -1, x_3 = x_4 = 0, x_5 = x_6 = 1$$

$i = 1, \ldots, 6$, the prior for the parameter α is $N(a, 1)$, the prior for the parameter β is $N(b, 1)$, both parameters are independent.

(a) Give the least squares estimators of α and β.

(b) Derive the posterior distribution of α and β given $Y = (y_1, \ldots, y_n)$.

(c) Derive the MAP estimators of α and β.

(d) Derive the marginal distribution of $Y = (y_1, \ldots, y_n)$.

8. Let us study the movement of a little mouse in a bounded flat area. Its position $x(t) = (x_1(t), x_2(t))$ is observed at time points $t \in (t_1, \ldots, t_n)$. We assume that

$$x(t) = x(t-1) + R_t \delta_t + \varepsilon_t$$

where $\delta_t = (\cos(\phi_t), \sin(\phi_t))$ with $\|\delta_t\| = 1$ describe the directions and R_t are the distances and ε_t are independent normal distributed random variables with variance σ^2. The mouse avoids the boundary. The following toy model is proposed:

$$\phi_t \sim 2\pi \, beta(\alpha, \beta), R_t \sim U(0, d).$$

When the mouse feels (:-)) that the next step will bring her too near to the border, she forgets her plan and makes new decisions until she can continue her path without danger. The unknown parameter ϑ consists of α, β, d.

(a) The aim is to generate pathes $z(t), t \in (t_1, \ldots, t_n)$ which are nearly the same as the observed $x(t), t \in (t_1, \ldots, t_n)$. Propose a criterion that two pathes are similar.

(b) Given the parameter ϑ how you would generate a path $z(t), t \in (t_1, \ldots, t_n)$? Give the main R commands.

(c) Write down the steps of an ABC algorithm for generating a sample of ϑ, distributed along the posterior distribution.

Appendix

$$Cov(X,Y) = E_Z(Cov((X,Y)|Z) + Cov_Z(E(X|Z), E(Y|Z))$$

$$\begin{pmatrix} a+1 & 1 & \cdots & 1 \\ 1 & a+1 & \vdots & \vdots \\ \vdots & \cdots & \ddots & 1 \\ 1 & \cdots & 1 & a+1 \end{pmatrix}^{-1} = (a\mathbf{I}_n + \mathbf{1}\mathbf{1}^T)^{-1} = \tfrac{1}{a}\mathbf{I}_n - \tfrac{1}{a(a+n)}\mathbf{1}\mathbf{1}^T$$

$$\left(\mathbf{I}_n + \mathbf{X}\mathbf{X}^T\right)^{-1} = \mathbf{I}_n - 2\mathbf{X}\mathbf{X}^T, \text{for } \mathbf{X}^T\mathbf{X} = \mathbf{I}_p$$

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right)$$

then

$$X_1|X_2 = x_2 \sim N\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$$