MSE | MASTER OF SCIENCE IN ENGINEERING

Final Examination 2018/19

Advanced Statistical Data Analysis

Thursday, 30th January 2019, 09:15 – 11:15

Prof. Dr. Andreas Ruckstuhl: Problems 1 and 2 (32 points)
Prof. Dr. Lin Himmelmann: Problems 3 to 8 (24 points)

Last name, first name: Musterlösung

UAS, part of UAS: _____

Achieved points: 24

General information:

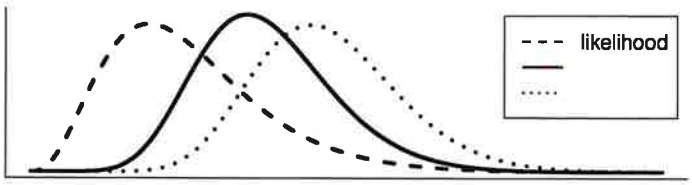
1. Write your name on the first page and on the top of every page.
2. The questions may be answered in German or in English.
3. Please answer directly on the question sheet. You may also use the back side.
4. If you need supplementary sheets, please use a separate one for every question. Write your name on every supplementary sheet.
5. Material allowed during the exam: calculator, manuscripts and books, no computer, no cell phone, no iPad, etc.
6. No question concerning the problems will be answered during the exam. If you don't understand a problem, make an assumption and explain it in your solution. It will be considered by the grader.
7. Communication with others during the exam is forbidden. Mobile phones must be turned off and dropped off at the front desk.
8. Do not write in red. This color is reserved for grading.
9. Portions of answers that have been crossed out won't be considered, even if the deleted part is correct.

Good luck!

Part 2: Problems 3-8 (24 points)

Problem 3 (1 point)

In the picture below are plotted the densities of the prior, posterior and likelihood. The dashed graph on the left refers to the likelihood.



Tick the correct solution(s):

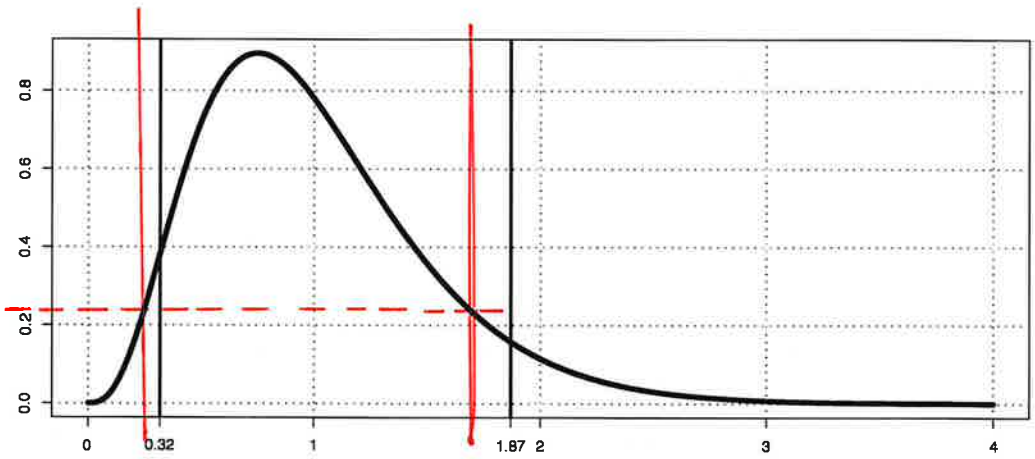
☐ The solid line must refer to the prior.

☒ The dotted line must refer to the prior.

☐ Both, the solid line as well as the dotted line could refer to the prior.

Problem 4 (2 points)

In the graph below is given a posterior density with a 90%-credible intervall $[0.32, 1.87]$.



Draw vertical lines at the approximately estimated bounds of the 90%-HDI (i.e. the 90%-credible interval with the highest posterior density) in the upper graph.

Problem 5 (8 points)

You are estimating the fraction θ of apple trees in Switzerland with a certain fungal disease. For this you selected 300 randomly sampled trees in Eastern Switzerland. 19 of them were infected with the fungal disease. Furthermore you selected 200 randomly sampled trees in Western Switzerland. 11 of them were infected with the fungal disease.

Based on observations of the previous years you assume a Beta distributed prior, where in expectation 5% of the trees are infected. The standard deviation is 0.05.

a) (3 points) Calculate the parameters for the Beta distributed prior.

$\frac{a}{a+b} = 0{,}05 \iff 20a = a+b \iff 19a = b$

$\frac{ab}{(a+b+1)(a+b)^2} = 0{,}05^2 \iff 400 \cdot 19a^2 = (20a+1)(20a)^2$

$\iff 20a+1 = 19 \iff a = \frac{18}{20} = \frac{9}{10} = 0{,}9$

$\Rightarrow b = \frac{171}{10} = 17{,}1$

b) (2 points) Specify the posterior distribution of θ.

$\theta | D \sim \text{Beta}(0{,}9 + 19 + 11,\ 17{,}1 + 200 + 300 - 19 - 11)$

$= \text{Beta}(30.9,\ 487.1)$

c) (1 point) Calculate the posterior mean of θ.

$\frac{30.9}{30.9 + 487.1} = 0.0597$

d) (1 point) Write out in propper R code: Compute a 90%-credible interval.

```
qbeta( c(0.05, 0.95), 30.9, 487.1)
```

e) (1 point) Write out in propper R code: Plot the posterior density.

```
curve( dbeta(x, 30.9, 487.1))
```

Problem 6 (4 points)

Consider you have 6 observations 3.1, 2.3, 5.4, 3.2, 5.8, and 4.2 from a normal distribution with known variance 2. Calculate the posterior distribution for the mean μ of the normal distribution, when you assume a normal distributed prior on μ, i.e. $\mu \sim Norm(3, 5^2)$.

$\sigma^2 = 2$

$3.1 + 2.3 + 5.4 + 3.2 + 5.8 + 4.2 = 24$

$$\mu | D \sim Norm\left(\left(\frac{3}{5^2} + \frac{24}{2}\right)\left(\frac{1}{5^2} + \frac{6}{2}\right)^{-1}, \left(\left(\frac{1}{5^2} + \frac{6}{2}\right)^{-\frac{1}{2}}\right)^2 \right)$$

$$= Norm\left(12.12 * (3.04)^{-1}, \left(\frac{1}{\sqrt{3.04}}\right)^2\right)$$

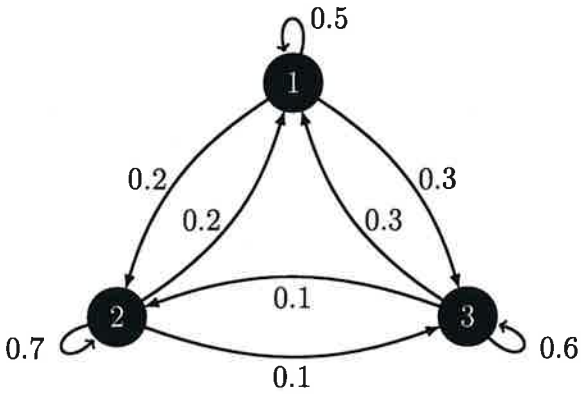
$$= Norm\left(3.987, 0.5735^2\right)$$

$$= Norm(3.987, 0.329)$$

$\frac{303}{76}$ $\frac{25}{76}$

Problem 7 (3 points)

Consider the following Markov model:



a) (1 point) Does this Markov model satisfy 'detailed balance'? Justify your answer!

for $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$

$\pi_i \cdot P_{i \to j} = \pi_j P_{j \to i}$ $\forall i,j = 1,2,3$, $i \neq j$

detailed balance holds!

b) (2 points) Is $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ a stationary distribution? Justify your answer!

$0.5 \cdot \frac{1}{3} + 0.2 \cdot \frac{1}{3} + 0.3 \cdot \frac{1}{3} = \frac{1}{3}$ ✓

$0.7 \cdot \frac{1}{3} + 0.2 \cdot \frac{1}{3} + 0.1 \cdot \frac{1}{3} = \frac{1}{3}$ ✓

$0.6 \cdot \frac{1}{3} + 0.3 \cdot \frac{1}{3} + 0.1 \cdot \frac{1}{3} = \frac{1}{3}$ ✓

yes, $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ is a stat. distribution.

Problem 8 (6 points)

Let $x = (x_1, \ldots, x_8) = (1.4, 3.2, -0.7, 2.6, 1.6, -1.2, 3.2, 3.7)$ be a vector of input values with corresponding responses $y = (y_1, \ldots, y_8) = (4.7, 7.1, 4.7, 5.7, 3.1, 2.3, 7.4, 5.0)$.

Consider a linear model, i.e. $y_i = mx_i + q + \epsilon_i$, with Gaussian noise $\epsilon_i \sim Norm(0, s^2)$. Assume a Gamma(2,4)-prior on s, a flat prior on q and a Gaussian prior with mean 0 and standard deviation 3 on m.

You've written the following script in R to estimate the posterior of m, q and s in a Bayesian analysis.

```
[01]  # Observed data:
[02]  obs_x = c(1.4, 3.2, -0.7, 2.6, 1.6, -1.2, 3.2, 3.7)
[03]  obs_y = c(4.7, 7.1, 4.7, 5.7, 3.1, 2.3, 7.4, 5.0)
[04]  # Function to calculate the (unnormalized) posterior density:
[05]  logStatDist = function(m_ , q_ , s_){
[06]        ...
[07]  }
[08]  # Choose random start values:
[09]  m = runif(1, 0, 1)
[10]  q = runif(1, 0, 1)
[11]  s = runif(1, 0, 1)
[12]  # Collect sampled values in these vectors:
[13]  m_sample = c( )
[14]  q_sample = c( )
[15]  s_sample = c( )
[16]  # Bayesian Data Analysis via MCMC:
[17]  for(i in 1:10000) {
[18]        m_prop = rnorm(m, 0, 0.02)
[19]        q_prop = rnorm(q, 0, 0.02)
[20]        s_prop = abs(rnorm(s, 0, 0.01))
[21]        R = exp(logStatDist(m_prop, q_prop, s_prop) - logStatDist(m,q, s))
[22]        u = runif(1,0,1)
[23]        if( u < R) {
[24]            m = m_prop
[25]            q = q_prop
[26]            s = s_prop
[27]        }
[28]        m_sample = c(m_sample, m)
[29]        q_sample = c(q_sample, q)
[30]        s_sample = c(s_sample, s)
[31]  }
```

a) (4 points) Write out the function logStatDist in line [05] from the R script in propper R code.

```
logStatDist = function(m_, q_, s_)
{
  sum( dnorm(m_ * obs_x + q_ - obs_y, 0, s_,
             log=TRUE))
  + dgamma(s_, 2, 4, log=T)
  + dnorm(m_, 0, 3, log=T)
}
# Since the prior on q is flat, you don't
  need to consider it in the code.
```

b) (2 points) After analysing the output, you found that the correlation of sampled values is too high. To avoid this, make a proposal on how to change the code to lower the acceptance rate.

Select proposed values from a large neighbour hood, for example:

```
[18]  m_prop = rnorm(m, 0, 0.05)
[19]  q_prop = rnorm(q, 0, 0.1)
[20]  s_prop = abs(rnorm(s, 0, 0.1))
```

END