Bayesian Data Analysis - Assignment 4

General information

- The recommended tool in this course is R (with the IDE R-Studio). You can download R **here** and R-Studio **here**. There are tons of tutorials, videos and introductions to R and R-Studio online. You can find some initial hints **here**.
- You can write the report with your preferred software, but the outline of the report should follow the instruction in the R markdown template that can be found **here**.
- Report all results in a single, **anonymous** *.pdf -file and return it to **peergrade.io**.
- Many of the exercises can be checked using the R package `markmyassignment`. Information on how to install and use the package can be found **here**.
- The course has its own R package with data and functionality to simplify coding. To install the package just run the following:

  1. `install.packages("devtools")`
  2. `devtools::install_github("avehtari/BDA_course_Aalto", subdir = "rpackage")`

- Many of the exercises can be checked automatically using the R package `markmyassignment`. Information on how to install and use the package can be found **here**.
- Additional self study exercises and solutions for each chapter in BDA3 can be found **here**.
- If you have any suggestions or improvements to the course material, please feel free to create an issue or submit a pull request to the public repository!!

Information on this assignment

This exercise is related to Chapters 3 and 10. The maximum amount of points from this assignment is 6.

Reading instructions: Chapters 3 and 10 in BDA3, see reading instructions **here** and **here**

Grading instructions: The grading will be done in peergrade. All grading questions and evaluations for exercise 4 can be found **here**

To use markmyassignment for this assignment, run the following code in R:

```
> library(markmyassignment)
> exercise_path <-
+   "https://github.com/avehtari/BDA_course_Aalto/blob/master/exercises/tests/ex4.yml"
> set_assignment(exercise_path)
```

---

1. **(Bioassay model and importance sampling)**. In this exercise, you will use a dose-response relation model that is used in Section 3.7 of the course book. The used likelihood is the same, but we will use a different prior distribution on the parameters α and β.

   a) Construct a bivariate normal distribution as prior distribution for (α, β). The marginal distributions are $\alpha \sim N(0, 2^2), \beta \sim N(10, 10^2)$ with correlation $\text{corr}(\alpha, \beta) = 0.5$. Report the mean and covariance of the bivariate normal distribution.
   Hint! The mean and covariance of the bivariate normal distribution are a length-2 vector and a 2×2 matrix. The elements of the covariance matrix can be computed using the relation of correlation and covariance.

   b) Implement a function in R for computing the **logarithm** of the density of the prior distribution in a) for arbitrary values of α and β. Below is an example of how the function should be named and work if you want to check them with `markmyassignment`.

   ```
   > alpha <- 3
   > beta <- 9
   > p_log_prior(alpha,beta)
   [1] -6.296435
   ```

   Hint! Use R function `dmvnorm` from the `mvtnorm` package. We use logarithms for better numerical accuracy in later questions.

   c) Implement a function in R for computing the **logarithm** of the density of the posterior for arbitrary values of α and β. Below is an example of how the function should be named and work if you want to check them with `markmyassignment`.

   ```
   > alpha <- 3
   > beta <- 9
   > p_log_posterior(alpha,beta)
   [1] -15.78798
   ```

   Hint! Equation (3.16) in the course book. The **logarithm** of the prior density was already implemented in b). For computing the **logarithm** of the likelihood, use the `bioassaylp` function from the `aaltobda` package. The data can be loaded with the R command `data("bioassay")`.
   Hint! Logarithm of the product of two densities is the sum of the log-densities, i.e. $\log ab = \log a + \log b$.

   d) Plot the posterior density in a grid of points ($\alpha \in [-4, 4]$, $\beta \in [-10, 30]$) using the `bioassay_posterior_density_plot` function from the `aaltobda` package. Internally, it uses the `p_log_posterior` function you implemented in c).

   e) Sample draws of α and β from the prior distribution from a). Compute the importance ratios (importance weights) for each draw when the target distribution is the posterior distribution. Normalize the weights so that they sum to 1. Below is a test example, the functions can also be tested with `markmyassignment`.

```
> alpha <- c(1.896, -3.6,  0.374, 0.964, -3.123, -1.581)
> beta <- c(24.76, 20.04, 6.15, 18.65, 8.16, 17.4)
> log_importance_weights(alpha, beta)
[1]  -8.95 -23.47  -6.02  -8.13 -16.61 -14.57
> normalized_importance_weights(alpha, beta)
[1] 0.045 0.000 0.852 0.103 0.000 0.000
```

Hint! Use R function `rmvnorm` from the `mvtnorm` package. Equation (10.3) in the course book.

f) Compute the posterior mean using importance sampling and draws from e), and report it.

```
> posterior_mean(alpha, beta)
[1] 0.503 8.275
```

g) Using the importance ratios, compute the effective sample size S_{eff} and report it. If S_{eff} is less than 1000, repeat e) and f) with more draws.

```
> S_eff(alpha, beta)
[1] 1.354
```

Hint! Equation (10.4) in the course book.

- **Note!** *BDA3 1st (2013) and 2nd (2014) printing have an error for* $\tilde{w}(\theta^s)$ *used in the effective sample size equation (10.4). The normalized weights equation should not have the multiplier S (the normalized weights should sum to one). Errata for the book* http://www.stat.columbia.edu/~gelman/book/errata_bda3.txt. *The later printings and slides have the correct equation.*

h) Use importance resampling without replacement to obtain a posterior sample of size 1000 of α and β and plot a scatterplot of the obtained posterior sample.

i) Using the posterior sample obtained via importance resampling, report an estimate for $p(\beta > 0|x, n, y)$, that is, the probability that the drug is harmful.

j) Using the posterior sample obtained via importance resampling, draw a histogram of the draws from the posterior distribution of the LD50 conditional on $\beta > 0$.
Hint! See Figure 3.4 and corresponding section in the course book.