# Take Home Exam

Chun-Biu Li
Felix Günther

This take home exam consists of 4 exercises. The report of the exam needs to be handed in separately by each student signed-up in Ladok in order to obtain the ECTS credits for the course. Your solution shall consist of the following:

- A written report as a PDF file containing solutions in the form of results, textual interpretations and graphs for the 4 exercises. The report must be completed independently. Plagiarism or other forms of cheating is a serious act. To underline this, the *signed confirmation sheet* must be submitted to declare that your work is made in accordance with the rules for written exams at Stockholm University (see course page).
- A file `<lastname>.R` containing the R code used to obtain all results and graphics contained in the report. Structured and well-documented code is important, e.g., each function should be preceded by a short text explaining what the input and output parameters are. Further code comments are to be made where needed and indentation should be used – see, e.g., Google's R Style Guide for further guidelines. Results are not to be discussed in the code – this is done in the report. As a trivial quality check: the command `source("<lastname>.R")` should run without errors for your code file.
- Deadline: **Sunday Oct. 31 2021 at 6:00pm**. The report has to be handed in as a bundle consisting of a) A scanned copy of your signed `Confirmation.pdf`, b) a PDF file `<lastname>.pdf` containing your report, c) the R file `<lastname.R>` and d) (in case of Sweave/knitr) `lastname.R[nw|md]` before the deadline. If you modified the original data or if your R code relies on external files (e.g., STAN models in text files), your bundle should contain these files as well (optimally as a ZIP file). All files are to be uploaded before the deadline to the Moodle drop-box on the course home page. Please note that there is a 10Mb file limit when uploading files. Delayed hand-ins are not accepted.

A total of 100 points can be reached for the answers in the report. Furthermore, up to 5 additional bonus points can be obtained, should your report and code be written with knitr. In this case please also attach the file `<lastname>.R[nw|md]` to your upload. Your final grade is determined by your sum of regular points and bonus points. Note: A penalty is imposed on reports longer than 28 pages.

Lycka till!

# Take Home Exam

Chun-Biu Li
Felix Günther

## Exercise 1 (23 points total)

This exercise is about comparing the results of Bayesian inference for different sampling schemes and priors.

a. (4 points) Suppose that one want to infer the probability of defection $\theta$ ($0 \leq \theta \leq 1$) for a product manufactured from a factory. From a random sample of 30 products, it is found that 3 of them are defected. Assume that the probability for a product to be defected is independent of the others. Use the uniform prior, find the posterior of $\theta$.

b. (6 points) Instead of fixing the number of samples as above, another sampling scheme is as follows. We keep on sampling the products randomly until 3 defected products are seen. It just happens that the 30th product sampled is the third defected one we found. Again using the uniform prior, find the posterior of $\theta$ and compare the results of Parts (a) and (b). *Hint*: The variable corresponds to the "data" in the likelihood function $p(data|\theta)$ is different from that of Part (a).

c. (5 points) Repeat Part (a) but using Jeffreys prior, find the prior and posterior as a function of $\theta$.

d. (4 points) Repeat Part (b) but using Jeffreys prior, find the prior and posterior as a function of $\theta$.

e. (4 points) Plot all four posteriors from Part (a)-(d) together and discuss how the design of the sampling scheme and the choice of prior can affect the results.

## Exercise 2 (27 points total)

This exercise contains some of the recommended exercises given in the class. Clear steps must be shown to receive full points. For Parts (a)-(c), suppose the likelihood $f(x|\theta, \sigma)$ is normal with mean $\theta$ and standard deviation $\sigma$:

a. (2 points) Suppose $\sigma$ is known, find the Jeffreys prior for $\theta$.

b. (2 points) Suppose $\theta$ is known, find the Jeffreys prior for $\sigma$.

c. (4 points) Suppose both $\theta$ and $\sigma$ are unknown, find the bivariate Jeffreys prior for $(\theta, \sigma)$. Discuss the difference from the two previous results.

d. (6 points) In class, we have shown for the univariate case that the Jeffreys prior $p(\theta) \propto [I(\theta)]^{1/2}$ is invariant to 1-1 transformation. Show that for the multivariate case, the Jeffreys prior $p(\theta) \propto |\mathbf{I}(\theta)|^{1/2}$ is invariant under parameter transformation, where $|\cdot|$ denotes the determinant, and $\mathbf{I}(\theta)$ is the expected Fisher information matrix. Notation of this part is the same as those in the course book.

e. (6 points) From the paper by Tierney and Kadane 1986 on Laplace approximations, derive Eq. (A.1) by showing that the leading correction term has the form $a/n$ with the coefficient $a$ as shown in the paper, where $n$ is the data size. Note: you do not need to work out the higher order terms, i.e., $b/n^2$ and $O(n^{-3})$.

f. (3 points) In the paper Sunnaker *et al.*, "Approximate Bayesian Computation", PLOS Comput. Biol., 9:e1002803 (2013), a nice conceptual overview (Fig. 1) was provided in the review concerning parameter estimation in ABC. Draw a similar conceptual overview for model comparison using ABC (refer to the section "Model Comparison with ABC").

g. (4 points) List up 3 concepts mentioned in the ABC paper that you are not familiar with and would like to learn more. For each of the 3 concepts, specify a potential reference where you can learn about the concept and briefly explain why the reference is picked. Word limit: $< 120$ words

# Exercise 3 (32 points total)

This exercise focuses on Bayesian linear regression, convergence diagnostics, model diagnostics and extensions of basic linear regression in STAN. It is based on an artificial dataset containing information on the growth/weight of 73 young leopards (younger than 16 month). For each animal it contains information on its age (in month) and weight (in kg). The dataset `growth.csv` can be downloaded from the moodle page of the course.

**Note:** Please document any warnings you encounter regarding the MCMC convergence diagnostics of your STAN model estimates. However, you do not necessarily have to optimise your code so that every warning disappears, usually the default settings (4 chains, 2000 iterations) should yield sufficient exploration of the posterior. Note, however, that warnings can also indicate errors or problems in your implementation (e.g., w.r.t coding, scaling, or chosen prior distributions). In case of major problems, discuss possible causes or possible solutions!

a. (6 points) Estimate an ordinary Bayesian linear regression model with outcome variable weight and covariate age using vague/non-informative priors for the unknown parameters using STAN. Investigate and interpret the posterior distribution of the regression coefficients. Comment briefly on the convergence of MCMC optimisation with reference to appropriate diagnostics.

b. (2 points) Calculate the standardised residuals of the regression model using the conditional expectation $E(Y|X = x, \hat{\theta})$, i.e. using a point estimate (e.g. the posterior median) of the parameters and neglecting the posterior uncertainty. Plot these residuals against age. Which assumption of ordinary linear regression is possibly violated?

c. (4 points) Think about a posterior predictive check that could be used to analyse deviation from the assumption in more detail and describe the approach as a pseudo-algorithm!

Read the case study on estimating non-linear associations via spline regression in STAN: https://mc-stan.org/users/documentation/case-studies/splines_in_stan.html.

d. (4 points) Investigate the STAN model `spline_reg.stan` from the webpage that can be used to fit a linear regression with non-linear effects via splines. Describe each component and its functionality shortly, e.g., by adding short comments to each row of the model. Provide a bit more detail on the two objects `des_mat` and `des_mat_exp_grid`. Why do they have the specified dimensions and what is contained in the individual entries of these arrays? *Hint: the latter of the two is only used to compute the model-based posterior expectation $E(Y|x, \theta)$ on a regular (or arbitrary) grid of $X$ (grid with 100 steps). It is not directly necessary for model fitting.*

e. (6 points) Use the `bs` function of the `splines` package in R to create a spline basis for the observed age covariate of the leopard data using 12 evenly-spaced knots in the range 0-16 month and third degree basis functions (*cubic splines*). Use this spline-basis (and a corresponding one for `des_mat_exp_grid`) to estimate the association of weight and age based on the `spline_reg.stan` model. Plot the estimated posterior distribution of $E(Y|\theta, X)$ and compare it to the scatterplot of the observed data. Interpret and/or criticise the results.

f. (10 points) Explain how prior distributions can be used to obtain a smoother estimation of the non-linear association of age and weight in the leopard data. Extend the `spline_reg.stan` model correspondingly and fit it to the data using the same spline basis as in part e.).

   Compare estimates of the posterior predictive distribution for weight on a grid of age values based on all three models. Comment on your results.

# Exercise 4 (18 points total)

This exercise is focused on Bayesian logistic regression based on data of a toxicity test as part of a drug-development process. The data consists of 4 batches of 5 animals that are administered different doses of the drug. The data comes in form of

$$(x_i, n_i, y_i),\ i = 1, \ldots, 4,$$

where $x_i$ represents the log-dose level in batch $i$, $n_i = 5$ the number of animals per batch, and $y_i$ the number of tumor developments.

We are interested in modeling the dose-response relation (occurrence of tumors due to the toxicity of too high doses) based on logistic regression:

$$y_i|\pi_i \sim Bin(n_i, \pi_i),$$

with $\pi_i = 1/(1 + exp(-(\alpha + \beta * x_i)))$. We will do this in a Bayesian setting (with prior-information on the parameters $\theta = (\alpha, \beta)$), and your task will be to implement a Metropolis-Hastings algorithm for sampling from the posterior distribution.

The dataset *toxic.csv* is available on the moodle page.

a. (4 points) Derive the likelihood and log-likelihood of the observed data as a function of $\theta = (\alpha, \beta)$.

b. (4 points) Implement the log-likelihood of the observed data (for arbitrary parameters $\theta = (\alpha, \beta)$) as well as the un-normalized posterior distribution of the parameters on the log-scale as a function `unnorm_post = function(alpha, beta, x, y, n)` using a multivariate normal prior distribution:

$$(\alpha, \beta) \sim N(\mu_0, \Sigma_0),\ \mu_0 = (0, 10)^t,\ \Sigma_0 = \begin{bmatrix} 4 & 12 \\ 12 & 100 \end{bmatrix}.$$

Plot a contour plot of the un-normalized posterior in the range $\alpha \in [-2.5.5]$, $\beta \in [-1, 30]$.

**Hint:** You can use the functions `dbinom` and `dmvnorm` from the `mvtnorm` package as basis for constructing the log-likelihood and log-posterior. Note, that the unnormalized log-posterior just corresponds to $\sum_{i=1}^{k} \text{log-Lik}(\theta|\text{data}) + \text{log-prior}(\theta)$.

c. (6 points) Implement a Metropolis-Hastings sampler to sample from the posterior distribution using a simple multivariate normal proposal distribution

$$(\alpha_t, \beta_t)|(\alpha_{t-1}, \beta_{t-1}) \sim N((\alpha_{t-1}, \beta_{t-1})^t, \Sigma_p),\ \ \Sigma_t = \begin{bmatrix} \sigma_\alpha^2 & 0 \\ 0 & \sigma_\beta^2 \end{bmatrix}.$$

You can start with $\sigma_\alpha^2 = 1$ and $\sigma_\beta^2 = 5$ and try some additional values for better acceptance rates in part d) (no need to tune for optimal choices). Start with implementing a function that evaluates the ratio of the un-normalized posterior for different parameter values $\theta_t$ and $\theta_{t-1}$. Note that $p_1/p_0 = exp(log(p_1) - log(p_0))$. You can check your function by verifying that the ratio of the posterior at $\theta_{prop} = (2, 10)$ and $\theta_{old} = (2, 20)$ is 2.5097. At $\theta_{prop} = (0, 5)$ and $\theta_{old} = (0, 6)$ the ratio equals 0.9275.

d. (4 points) Use your algorithm from part c) to sample 10000 draws from the posterior distribution. Start with $(\alpha_0 = -2.5, \beta_0 = 0)$. Plot traceplots of the draws and discuss the idea of a warm-up phase. Investigate the marginal posterior distributions of $\alpha$ and $\beta$ and interpret your results. Plot (a random sample) of the bivariate draws of your MCMC algorithm to the contour plot of section b).