Bayesian data analysis, fall 2024
Course exam

This exam consists of four tasks, 6 points each. Write your name on each answer sheet. No calculators are needed, just pen and paper. Answer conscisely, always justifying your claims.

Task 1: Answer the following questions briefly (6 points in total)

a) What are posterior intervals?
b) What is the key motivation and workflow when doing cross-validation for a set candidate models?
c) Which elements of *Markov chain Monte carlo sampler* (for a fixed model and priors) affect its speed of convergence and how?
d) Explain the concept of *directed acyclic graph* using a drawn example.
e) Outline the idea behind computing WAIC (or AIC) and explain *how they can and cannot be used.*
f) Why should you run several MCMC chains when sampling from a posterior of interest?

Task 2. (6 points in total)

a) Explain the general idea of a hierarchical model, and why and how hierarchical models can be advantageous. Outline a concrete example in which such model could be appropriate, and define the model for that situation, either using mathematical notation or e.g. stan pseudocode, both with enough explanation.
b) In 1976 George Box stated: *All models are wrong, but some are useful.* Now, 48 years later, your task is to answer the following (within the context of the course): how can models go wrong? How do we construct as good as possible models? How can models be useful?

Task 3. (6 points in total; Appendix involved)

$p(\alpha)$
$p(\beta)$

Let's examine again data on whitefish larvae. Each observation contains the number of sampled larvae y_i in a sample of volume of z_i water. Let's assume we have data on environmental covariates $x_i = x_{i,1}, \ldots x_{i,5}$, from each sample location *i*, and the interest is to study their effects on the larvae density. For this we use a generalized linear model:

$\mu_i \sim$
$\sigma_i \sim$
$y_i \sim p(\mu_i, \sigma_i)$

$$y_i \sim Poisson(z_i * \lambda_i)$$

$$\log(\lambda_i) = a + b_1 x_{i1} + \ldots + b_5 x_{i5}$$

where $Poisson(z_i * \lambda_i)$ denotes for poisson distribution with mean parameter $(z_i * \lambda_i)$. The other parameters are intercept *a* and linear weights $b_1, \ldots, b_5$. Notice also that z_i is an observed covariate, but it is assumed to have a multiplicative effect directly on λ_i.

As an attachment, you'll have R output for the model above. Based on the assignment and the shown results answer the following questions:

a) Define non-informative priors for the model parameters.
b) Explain why you need to use log-link function in this model.
c) Write down pseudo-STAN-code for this model.
d) Is autocorrelation a problem here?
e) Explain step by step, using actual distributions, how you can predict *replicate data* from the posterior predictive distribution $p(y^{rep}|y)$ based on the given variables in the R outputs.
f) How good is the chosen model based on the posterior predictive check?

$P(\text{params} \mid y) \propto P(y \mid \text{params})\, p(\text{priors}) \ldots$

$\Rightarrow p(y^{rep} \mid y) \propto p\, y \mid$

Task 4. (6 points in total)

Let's consider four different models fit to one observation, which is: $y=0$. The models differ in the distributions assigned to the likelihood and to the prior. The models outlined given below, where first the assumed likelihood is given and then the prior for the model parameter (here only μ).

Model NN:

$$y \sim Normal(\mu,1)$$
$$\mu \sim Normal(10,1)$$

Model NT:

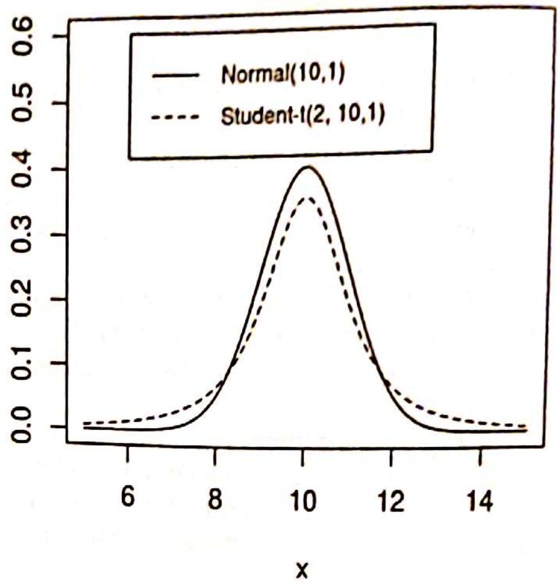
$$y \sim Normal(\mu,1)$$
$$\mu \sim Student(2,10,1)$$

Model TN:

$$y \sim Student(2,\mu,1)$$
$$\mu \sim Normal(10,1)$$

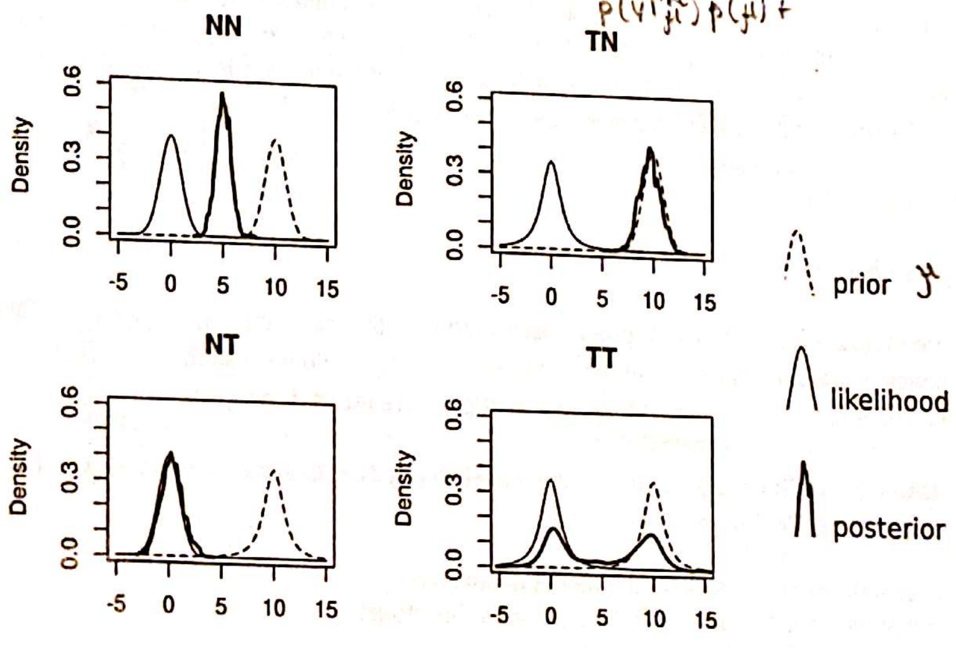
Model TT:

$$y \sim Student(2,\mu,1)$$
$$\mu \sim Student(2,10,1)$$



Here $Student(\nu,\mu,\sigma)$ refers to student's t distribution with ν degrees of freedom, mean μ and scale σ. Examples of relevant Normal and Students t distributions are plotted above in right. What happens in each model fit is shown below, where the names of the fitted model are given in titles.

$p(\mu|y) = \frac{p(y|\mu)\, p(\mu)\; N(10,1)}{p(y|\mu)\, p(\mu) +}$



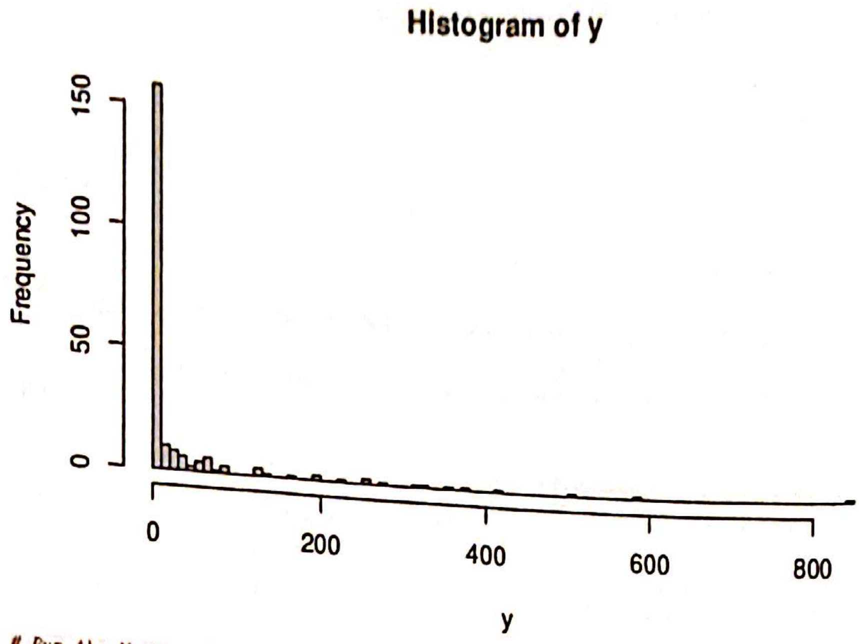
Your task is to:

a) Notice and explain the differences between the shown posterior distributions, and contemplate on the relevancy of each posterior.

b) Explain what sort of considerations are important when selecting prior distributions for different kinds of statistical modeling tasks.

Appendix: R-code and output for the white fish larvae model:

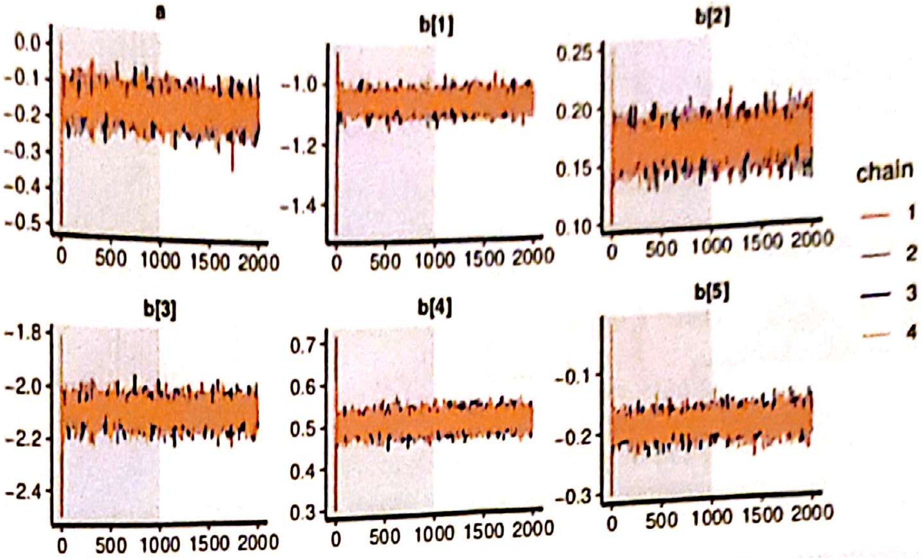
```
# Load the needed libraries into R
library(ggplot2)
library(StanHeaders)
library(rstan)

## rstan (Version 2.19.2, GitRev: 2e1f913d3ca3)

## For execution on a local, multicore CPU with excess RAM we recommend calling
## options(mc.cores = parallel::detectCores()).
## To avoid recompilation of unchanged Stan programs, we recommend calling
## rstan_options(auto_write = TRUE)

options(mc.cores = parallel::detectCores())
rstan_options(auto_write = TRUE)

# Load the data and put it into a list format
whitefish.dat = read.table("data_whitefish_2.csv", header=TRUE, sep=",")

# The columns are:
#    VOLUME     = volume of water sampled
#    WHISUM     = number of whitefish larvae caught
#    DEPTH      = depth
#    FE300ME    = openness
#    DIS_SAND   = distance to sandy shore
#    ICEWIN10   = length of ice cover season
#    SHAREA     = Amount of shallow area within "nearby" sampling site

y = whitefish.dat$WHISUM              # Species observations
z = whitefish.dat$VOLUME              # Sampling effort
x = as.matrix(whitefish.dat[,3:7])    # environmental covariates
N <- nrow(x)

# standardize covariates and use them in the analysis
xm = as.vector(colMeans(x))
xsd = apply(x, 2, sd)
x.st = t(apply(x, 1, "-", xm))
x.st = t(apply(x.st, 1, "/", xsd))

data <- list ("N"=N, "y"=y, "x"=x.st,  "z"=z)

hist(y,breaks=100,col='grey')
```



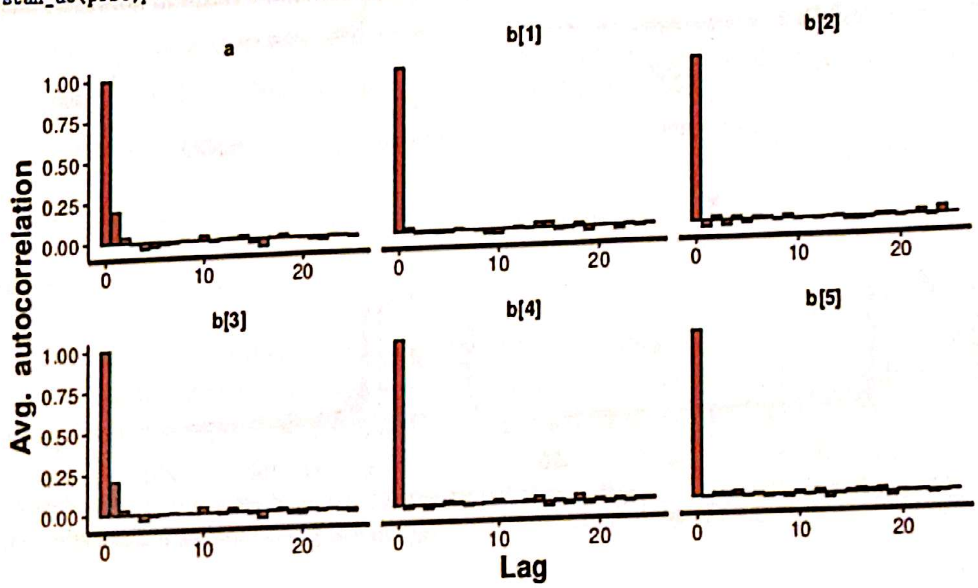

```
# Run the Markov chain sampling with Stan:

init1 <- list (a = -0.5, b = c(-1.5,0.1,-2.5,0.3,-0.3))
init2 <- list (a = -0.2, b = c(-1.2,0.15,-2.2,0.4,-0.2))
init3 <- list (a = -0.1, b = c(-1.1,0.2,-2,0.5,-0.1))
init4 <- list (a = 0, b = c(-1,0.25,-1.8,0.6,0))
inits <- list(init1, init2, init3, init4)

post=stan(file="Model_exam_30_1.stan",data=data,init=inits,warmup=1000,iter=2000,chains=4,thin=1)

summary(post,pars=c("a","b"))$summary
```

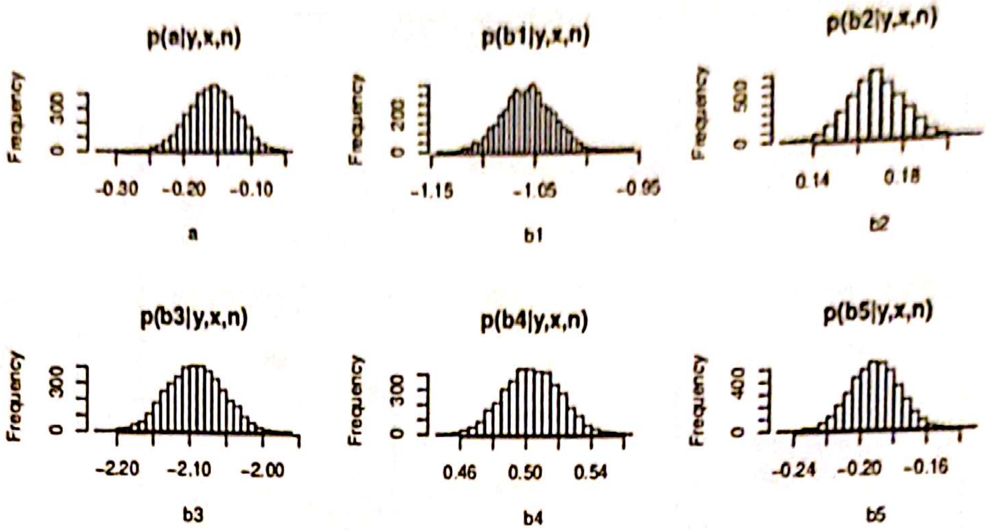
```
##            mean      se_mean         sd       2.5%        25%        50%
## a    -0.1569260 0.0006622016 0.03469168 -0.2249481 -0.1804342 -0.1567461
## b[1] -1.0570161 0.0003923832 0.02436689 -1.1053331 -1.0734495 -1.0568538
## b[2]  0.1677277 0.0001817599 0.01199165  0.1443361  0.1597798  0.1676834
## b[3] -2.0920983 0.0007329139 0.03818913 -2.1662315 -2.1187595 -2.0922932
## b[4]  0.5038883 0.0002690715 0.01737003  0.4702294  0.4918811  0.5037608
## b[5] -0.1915185 0.0002435660 0.01462375 -0.2202220 -0.2017187 -0.1915994
##             75%       97.5%    n_eff      Rhat
## a    -0.1335433 -0.09029583 2744.544 1.0004164
## b[1] -1.0405266 -1.01007260 3856.378 1.0006783
## b[2]  0.1756600  0.19114276 4352.731 1.0011819
## b[3] -2.0656676 -2.01852499 2715.024 1.0006699
## b[4]  0.5156833  0.53763417 4167.406 0.9995436
## b[5] -0.1814122 -0.16300636 3604.825 1.0002106
```

```
plot(post, plotfun= "trace", pars=c("a","b"), inc_warmup = TRUE)
```




```
samples <- as.matrix(post,pars=c("a","b"))
a <- samples[,1]
b1 <- samples[,2]
b2 <- samples[,3]
b3 <- samples[,4]
b4 <- samples[,5]
b5 <- samples[,6]

stan_ac(post,pars=c("a","b"),inc_warmup = FALSE, lags = 25)
```





#Conduct a visual posterior predictive check for the data. Sample replicates of the data set and plot the histogram for a couple of replicate data set.

```
par(mfrow=c(2,2))
for (i in 1:4){
  ind=floor( length(a)*runif(1))
  hist(rep_samples[ind,],breaks=100,main="",xlab="y replicate",col="grey")
}
```