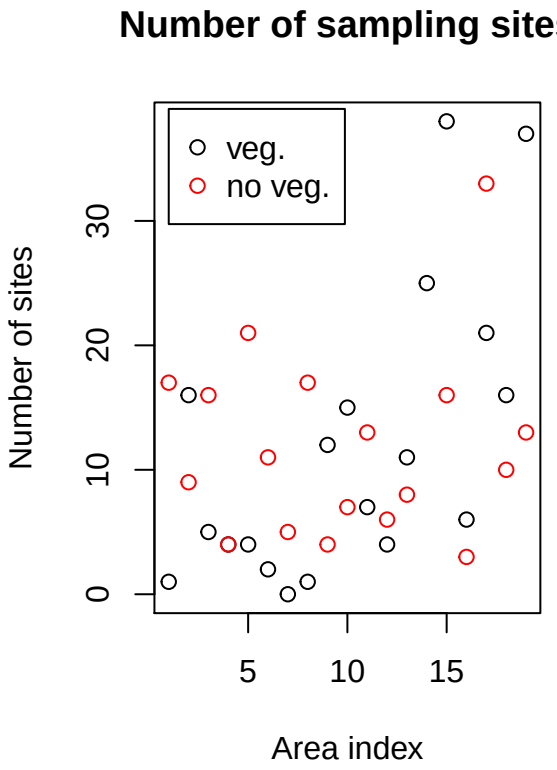
Effect of bottom coverage to larval presence

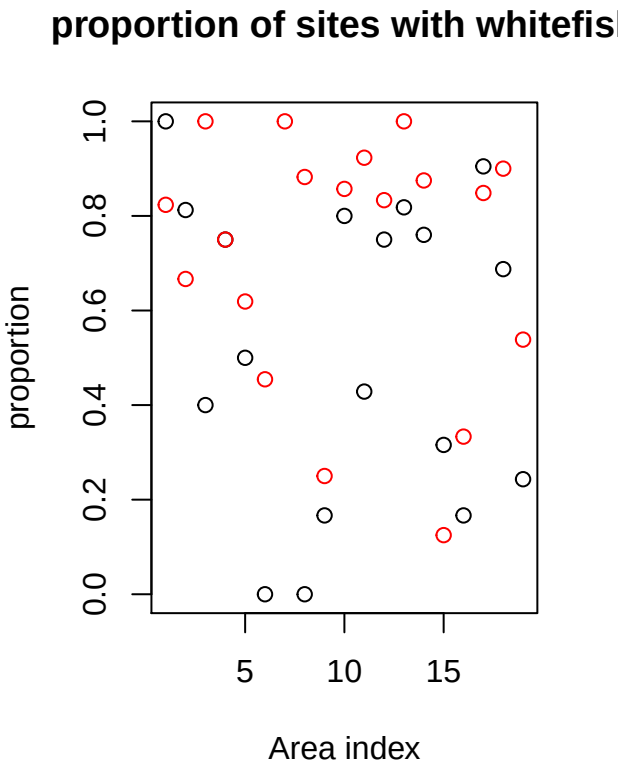
Week4-ex2, solution

In this exercise, we continue the analysis of the white fish larval areas (week 2, exercise 3). We are again interested in analysing whether or not bottom vegetation affects white fish larvae occurrence probability. However, instead of having a common probability of presence parameter across the Gulf of Bothnia, we expand the model so that it allows the probability of presence to vary between sampling areas. This modification to the model encodes an assumption that some areas may be more preferable to white fish than others.

Let's first explore the data a bit more.

```
# Read the data
data = read.csv("white_fishes_data.csv")
# Form a data table for sites without bottom vegetation
y.noveg = table(data$AREANAME[data$BOTTOMCOV==0], data$WHIBIN[data$BOTTOMCOV==0])
colnames(y.noveg) <- c("y=0","y=1")
N.noveg = rowSums(y.noveg)
# Form a data table for sites with bottom vegetation
y.veg = table(data$AREANAME[data$BOTTOMCOV==1], data$WHIBIN[data$BOTTOMCOV==1])
colnames(y.veg) <- c("y=0","y=1")
N.veg = rowSums(y.veg)

par(mfrow=c(1,2))
plot(N.veg, main="Number of sampling sites", xlab="Area index", ylab="Number of sites")
points(N.noveg, col="red")
legend(1, 39, c("veg.","no veg."),col=c("black","red"), pch=1, cex=1, box.lty=1)
plot(y.veg[,2]/N.veg, main="proportion of sites with whitefish", xlab="Area index", ylab="proportion")
points(y.noveg[,2]/N.noveg, col="red")
```






```
print(y.veg)
```

```
##
##                  y=0 y=1
##   Bjuroklubb       0   1
##   Bygdea           3  13
##   Haaparanta       3   2
##   Hailuoto         1   3
##   Harnosand        2   2
##   Hornslandet      2   0
##   Kalajoki         0   0
##   Lohtaja          1   0
##   Luvia           10   2
##   Mikkelinsaaret   3  12
##   Mjolefjarden     4   3
##   Nordingra        1   3
##   Pietarsaari      2   9
##   Pitea            6  19
##   Pori            26  12
##   Siipyy           5   1
##   Storsand         2  19
##   Tore             5  11
##   Vaasa           28   9
```

The first figure above shows the number of sampling sites for each of the 19 study areas and both bottom vegetation types (with and without). The second figure shows the proportion of the sites with white fish larvae within each area and bottom vegetation type. It is rather evident that there is considerable variation in the sample proportions of the second figure. However, we would want to know how much of this is actually due to varying probability of presence vs. pure chance. Note also, that there are no sampling sites in Kalajoki (sampling area number 7 below) with vegetation cover. Hence, we have missing data there.

```
N.veg
```

```
##     Bjuroklubb         Bygdea     Haaparanta       Hailuoto      Harnosand
##              1             16              5              4              4
##    Hornslandet       Kalajoki        Lohtaja         Luvia Mikkelinsaaret
##              2              0              1             12             15
##   Mjolefjarden      Nordingra    Pietarsaari         Pitea           Pori
##              7              4             11             25             38
##         Siipyy       Storsand          Tore         Vaasa
##              6             21             16             37
```

We will denote by $\theta_{i,c}$ the probability that white fish larvae are present in area i at sites with $(c = 1)$ or without $(c = 0)$ bottom vegetation. The data will be denoted by $y_{i,c}$ and $N_{i,c}$ where the former denotes the number of sites with white fish larvae and the latter the total number of sites inside an area i with $(c = 1)$ or without $(c = 0)$ bottom vegetation. We will now implement the following model

$$
\begin{aligned}
y_{i,c} &\sim \text{Binom}(\theta_{i,c}, N_{i,c}) \\
\theta_{i,c} &\sim \text{Beta}(\mu_c s_c, s_c - \mu_c s_c) \\
\mu_c &\sim \text{Unif}(0, 1) \\
s_c &\sim \text{log-}N(4, 4).
\end{aligned}
$$

where μ_c is the prior mean of $\theta_{i,c}$ and s_c governs the uncertainty about it. The parametrization of log-Gaussian distribution $s_c \sim \text{log-}N(m, \sigma^2)$ is such that $E[\log(s_c)] = m$ and $Var[\log(s_c)] = \sigma^2$

1. Implement the model in Stan and sample from the posterior for the parameters. Check for convergence for all parameters, and examine what is the autocorrelation for s_c, μ_c and few $\theta_{i,c}$. Visualize the posterior for μ_c, s_c and $\theta_{i,c}, i = 1, \ldots, 19$.
2. Visualize also the posterior distributions of $\Delta\mu = \mu_0 - \mu_1$ and $\phi_i = \theta_{i,0} - \theta_{i,1}$ for each area $i = 1, \ldots, 19$.
3. Sample from the posterior predictive distribution of outcome $\tilde{y}_{19,c}$ of a new sampling with $\tilde{N}_{19} = 10$ in the sampling area $i = 19$ for both vegetated and non-vegetated sites. Visualize the resulting posterior samples as well as the posterior distribution for $\tilde{y}_{19,0} - \tilde{y}_{19,1}$.
4. Sample from the posterior predictive distribution of outcome $\tilde{y}_{20,c}$ of a new sampling with $\tilde{N}_{20} = 10$ in a new sampling area $i = 20$ (an area from where we don't have data yet) within the Gulf of Bothnia. Do this for both vegetated and non-vegetated sites. Visualize the resulting posterior samples as well as the posterior distribution for $\tilde{y}_{20,0} - \tilde{y}_{20,1}$.
5. The posterior distributions calculated in exercise 3 of week 2 correspond to the so called pooled estimate of θ_c. Discuss how does the posterior of the pooled θ_c differ from the population mean, μ_c, and from the individual $\theta_{i,c}$ in the hierarchical model? Which model seems more justified in your opinion and why?

Solution

1.

Load the needed libraries into R and write the Stan model

```
library(ggplot2)
```

```
## Warning: replacing previous import 'vctrs::data_frame' by 'tibble::data_frame'
## when loading 'dplyr'
```

```
library(StanHeaders)
library(rstan)
```

```
## rstan (Version 2.19.3, 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)
```

```
library(gridExtra)
library(see)
options(mc.cores = parallel::detectCores())
rstan_options(auto_write = TRUE)
set.seed(123)

whitefish.model = "data{
  int<lower=0> n;       // number of sampling areas
  int<lower=0> N[n];    // number of sites per area
  int<lower=0> y[n];    // number of sites with white fish per area
}
parameters{
  vector<lower=0, upper=1>[n] theta;
  real<lower=0, upper=1> mu;
  real<lower=0> s;
}
model{

  theta ~ beta(mu*s,s-mu*s);
  mu ~ uniform(0,1);
  s ~ lognormal(4,2);

  for( i in 1 : n ){
  if (N[i]>0)
    y[i] ~ binomial(N[i],theta[i]);
  }
}"
```

Next, we create our data lists and do the sampling with both models

```
data.noveg <- list ("n"=dim(y.noveg)[1],"N"=N.noveg, "y"=y.noveg[,2])   # no vegetation data
data.veg <- list ("n"=dim(y.veg)[1],"N"=N.veg, "y"=y.veg[,2])   # vegetated bottoms data
post.noveg=stan(model_code=whitefish.model,data=data.noveg,
                warmup=200,iter=600,chains=3,thin=5,control=list(adapt_delta=0.99))
```

```
## Warning: Bulk Effective Samples Size (ESS) is too low, indicating posterior means and medians may be
## Running the chains for more iterations may help. See
## http://mc-stan.org/misc/warnings.html#bulk-ess
```

```
## Warning: Tail Effective Samples Size (ESS) is too low, indicating posterior variances and tail quanti
## Running the chains for more iterations may help. See
## http://mc-stan.org/misc/warnings.html#tail-ess
```

```
post.veg=stan(model_code=whitefish.model,data=data.veg,
              warmup=200,iter=600,chains=3,thin=5,control=list(adapt_delta=0.99))
```

```
## Warning: Bulk Effective Samples Size (ESS) is too low, indicating posterior means and medians may be
## Running the chains for more iterations may help. See
## http://mc-stan.org/misc/warnings.html#bulk-ess
```

```
## Warning: Tail Effective Samples Size (ESS) is too low, indicating posterior variances and tail quanti
## Running the chains for more iterations may help. See
## http://mc-stan.org/misc/warnings.html#tail-ess
```

Let's examine convergence and autocorrelation for both model fits (no-vegation and vegetation)

```
print(post.noveg)
```

```
## Inference for Stan model: d9f1caa65185894edb47f58cacee61e5.
## 3 chains, each with iter=600; warmup=200; thin=5;
## post-warmup draws per chain=80, total post-warmup draws=240.
##
##              mean se_mean   sd    2.5%     25%     50%     75%   97.5% n_eff
## theta[1]     0.81    0.01 0.08    0.64    0.76    0.82    0.87    0.94   230
## theta[2]     0.68    0.01 0.13    0.42    0.60    0.68    0.77    0.89   273
## theta[3]     0.94    0.00 0.05    0.80    0.92    0.96    0.99    1.00   240
## theta[4]     0.75    0.01 0.15    0.39    0.66    0.78    0.86    0.95   175
## theta[5]     0.63    0.01 0.10    0.41    0.56    0.64    0.70    0.81   288
## theta[6]     0.51    0.01 0.14    0.24    0.40    0.52    0.59    0.79   261
## theta[7]     0.87    0.01 0.11    0.62    0.81    0.90    0.96    1.00   234
## theta[8]     0.85    0.01 0.07    0.70    0.81    0.87    0.91    0.96    79
## theta[9]     0.48    0.01 0.17    0.14    0.37    0.49    0.60    0.76   174
## theta[10]    0.80    0.01 0.12    0.53    0.73    0.81    0.90    0.97   131
## theta[11]    0.89    0.00 0.07    0.72    0.85    0.90    0.94    0.98   238
## theta[12]    0.77    0.01 0.13    0.50    0.68    0.78    0.87    0.96   227
## theta[13]    0.91    0.01 0.09    0.65    0.89    0.93    0.98    1.00   214
## theta[14]    0.86    0.00 0.04    0.77    0.84    0.86    0.89    0.93   237
## theta[15]    0.24    0.01 0.10    0.08    0.16    0.23    0.32    0.46   260
## theta[16]    0.56    0.01 0.19    0.20    0.42    0.58    0.70    0.87   248
## theta[17]    0.84    0.00 0.06    0.69    0.79    0.84    0.88    0.94   251
## theta[18]    0.86    0.01 0.09    0.65    0.82    0.87    0.93    0.98   244
## theta[19]    0.59    0.01 0.12    0.37    0.52    0.60    0.68    0.81   164
## mu           0.71    0.00 0.06    0.59    0.68    0.72    0.75    0.81   153
## s            4.14    0.13 1.92    1.47    2.74    3.80    5.20    8.79   222
## lp__      -158.63    0.34 4.39 -168.67 -161.42 -158.16 -155.27 -152.02   167
##           Rhat
## theta[1]  1.02
## theta[2]  0.99
## theta[3]  0.99
## theta[4]  1.01
## theta[5]  0.99
## theta[6]  1.01
## theta[7]  1.00
## theta[8]  1.03
## theta[9]  1.01
## theta[10] 1.01
## theta[11] 1.00
## theta[12] 0.99
## theta[13] 1.02
## theta[14] 1.00
## theta[15] 1.00
## theta[16] 1.00
## theta[17] 0.99
## theta[18] 1.00
## theta[19] 1.00
```

```
## mu        1.01
## s         1.00
## lp__      1.00
##
## Samples were drawn using NUTS(diag_e) at Mon Nov  2 11:57:10 2020.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

```
p1 = plot(post.noveg, plotfun= "trace", pars=c("s","mu"), inc_warmup = TRUE)
p2 = plot(post.noveg, plotfun= "trace", pars=c("theta[1]","theta[2]"), inc_warmup = TRUE)
p3 = stan_ac(post.noveg,c("s","mu"),inc_warmup = FALSE, lags = 25)
```

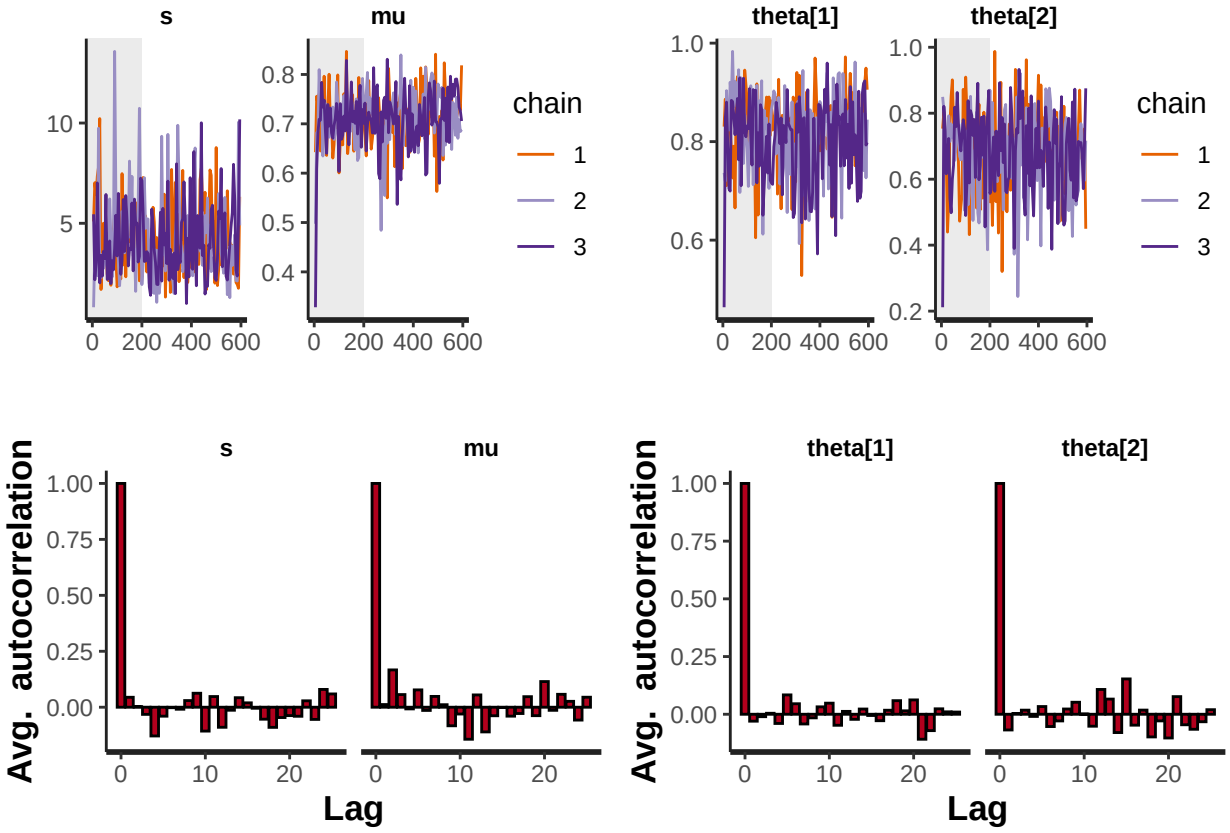
```
## Warning: Ignoring unknown parameters: fun.y
```

```
p4 = stan_ac(post.noveg,c("theta[1]","theta[2]"),inc_warmup = FALSE, lags = 25)
```

```
## Warning: Ignoring unknown parameters: fun.y
```

```
grid.arrange(p1,p2,p3,p4)
```

```
## No summary function supplied, defaulting to `mean_se()`
## No summary function supplied, defaulting to `mean_se()`
## No summary function supplied, defaulting to `mean_se()`
## No summary function supplied, defaulting to `mean_se()`
```




```
print(post.veg)
```

```
## Inference for Stan model: d9f1caa65185894edb47f58cacee61e5.
## 3 chains, each with iter=600; warmup=200; thin=5;
```

```
## post-warmup draws per chain=80, total post-warmup draws=240.
##
##              mean se_mean   sd    2.5%     25%     50%     75%   97.5% n_eff
## theta[1]     0.64    0.01 0.21    0.23    0.48    0.67    0.79    0.97   209
## theta[2]     0.75    0.01 0.09    0.56    0.69    0.76    0.82    0.90   226
## theta[3]     0.46    0.01 0.16    0.15    0.35    0.47    0.58    0.72   191
## theta[4]     0.65    0.01 0.16    0.33    0.55    0.66    0.77    0.92   259
## theta[5]     0.53    0.01 0.16    0.20    0.42    0.54    0.64    0.81   158
## theta[6]     0.36    0.01 0.19    0.05    0.21    0.36    0.49    0.74   261
## theta[7]     0.54    0.02 0.24    0.04    0.39    0.55    0.72    0.95   221
## theta[8]     0.43    0.01 0.21    0.06    0.28    0.44    0.60    0.81   263
## theta[9]     0.27    0.01 0.11    0.05    0.20    0.27    0.35    0.47   133
## theta[10]    0.74    0.01 0.10    0.52    0.69    0.75    0.82    0.90   271
## theta[11]    0.48    0.01 0.14    0.19    0.38    0.48    0.57    0.74   173
## theta[12]    0.63    0.01 0.16    0.33    0.52    0.63    0.75    0.90   190
## theta[13]    0.74    0.01 0.12    0.47    0.67    0.75    0.82    0.93   183
## theta[14]    0.73    0.01 0.09    0.54    0.67    0.73    0.79    0.88   233
## theta[15]    0.34    0.01 0.07    0.20    0.29    0.34    0.39    0.49   203
## theta[16]    0.31    0.01 0.15    0.07    0.20    0.30    0.42    0.62   233
## theta[17]    0.84    0.00 0.07    0.69    0.79    0.85    0.89    0.95   208
## theta[18]    0.65    0.01 0.11    0.44    0.57    0.66    0.72    0.85   223
## theta[19]    0.29    0.00 0.07    0.15    0.24    0.28    0.34    0.45   214
## mu           0.54    0.00 0.07    0.40    0.49    0.54    0.58    0.66   242
## s            4.73    0.18 2.64    1.57    2.91    4.08    5.59   11.08   209
## lp__      -158.65    0.29 3.93 -167.37 -161.40 -158.49 -155.89 -151.68   183
##           Rhat
## theta[1]  1.00
## theta[2]  0.99
## theta[3]  1.01
## theta[4]  0.99
## theta[5]  1.00
## theta[6]  0.99
## theta[7]  0.99
## theta[8]  1.00
## theta[9]  1.01
## theta[10] 1.02
## theta[11] 1.02
## theta[12] 1.00
## theta[13] 1.00
## theta[14] 1.01
## theta[15] 1.00
## theta[16] 1.00
## theta[17] 1.00
## theta[18] 1.01
## theta[19] 0.99
## mu        1.00
## s         0.99
## lp__      1.00
##
## Samples were drawn using NUTS(diag_e) at Mon Nov  2 11:57:12 2020.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

```
p1=plot(post.veg, plotfun= "trace", pars=c("s","mu"), inc_warmup = TRUE)
p2=plot(post.veg, plotfun= "trace", pars=c("theta[1]","theta[2]"), inc_warmup = TRUE)
p3=stan_ac(post.veg,c("s","mu"),inc_warmup = FALSE, lags = 25)
```

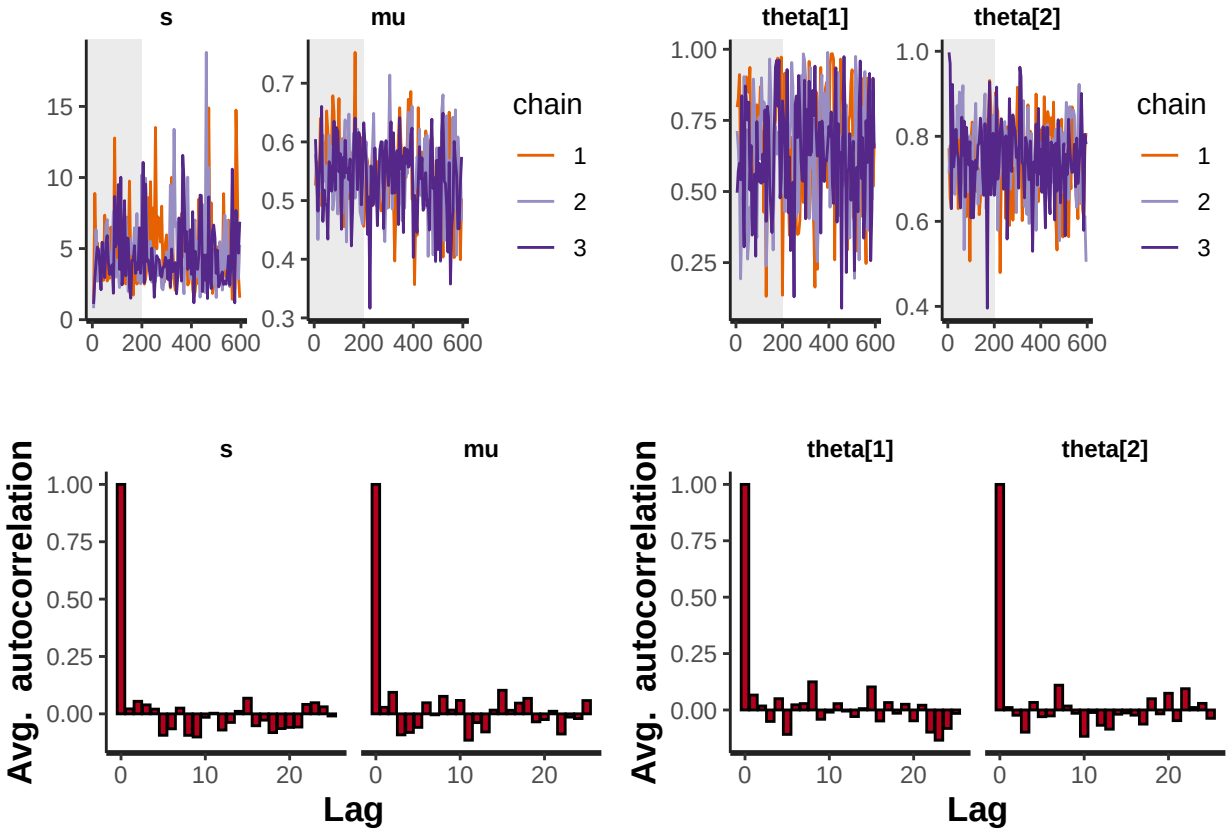
```
## Warning: Ignoring unknown parameters: fun.y
```

```
p4=stan_ac(post.veg,c("theta[1]","theta[2]"),inc_warmup = FALSE, lags = 25)
```

```
## Warning: Ignoring unknown parameters: fun.y
```

```
grid.arrange(p1,p2,p3,p4)
```

```
## No summary function supplied, defaulting to `mean_se()`
## No summary function supplied, defaulting to `mean_se()`
## No summary function supplied, defaulting to `mean_se()`
## No summary function supplied, defaulting to `mean_se()`
```



According to the Rhat summary and the plotted trace plots of sample chains the chains seem to have converged in both cases. The autocorrelation of the Markov chain samples is very small and, hence, not a problem.

Let's then visualize the posterior for μ, s and $\theta_i, i = 1, \ldots, 19$

```
library(ggplot2)
library(gridExtra)
library(see)
# create ggplots of the wanted parameters
p1 = plot(post.noveg, plotfun = "hist", pars = c("s"),bins=50) + ggtitle("no vegetation")
p2 = plot(post.noveg, pars = c("mu","theta"))
```

```
## ci_level: 0.8 (80% intervals)
## outer_level: 0.95 (95% intervals)
```

```
p3 = plot(post.veg, plotfun = "hist", pars = c("s"),bins=50) + ggtitle("vegetation")
p4 = plot(post.veg, pars = c("mu","theta"))
```

```
## ci_level: 0.8 (80% intervals)
## outer_level: 0.95 (95% intervals)
```

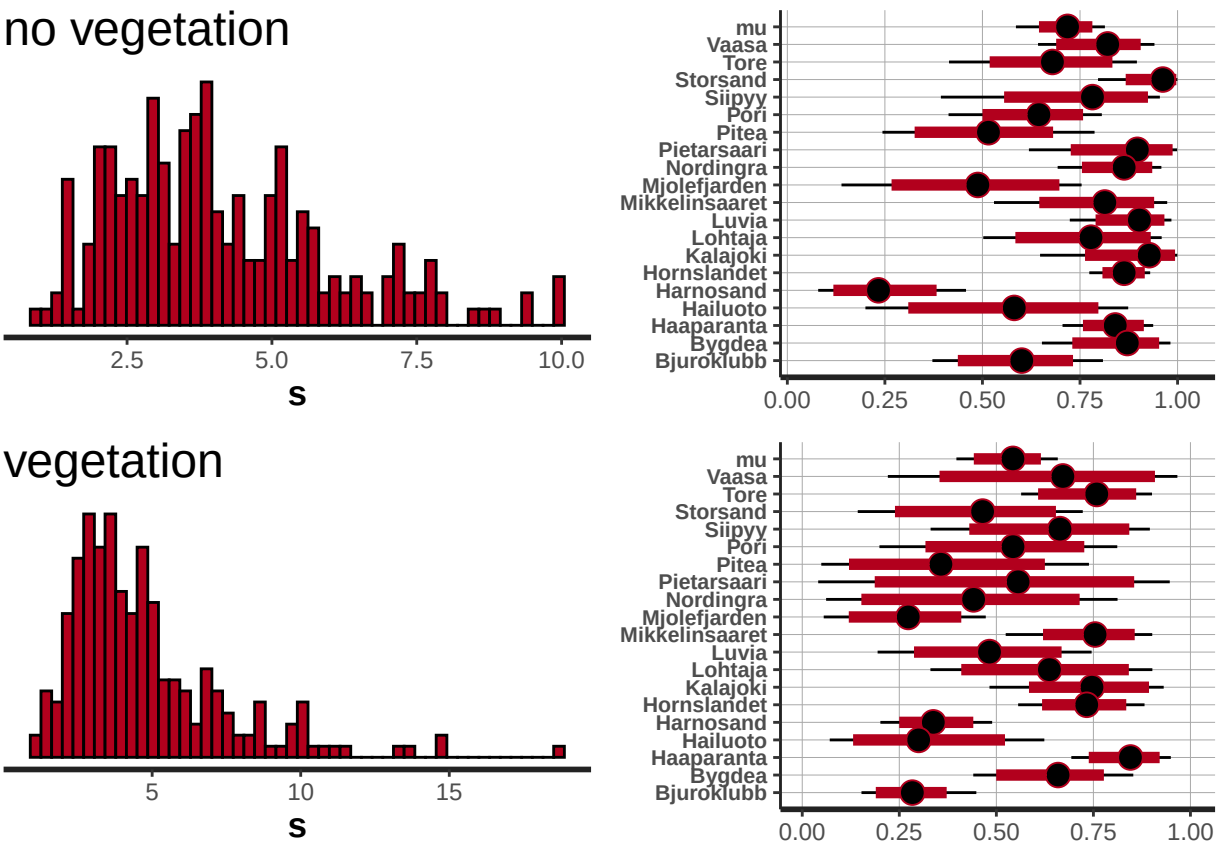
```
# Rename and scale the size of the y-ticks
p2 = p2 + scale_y_continuous(breaks=c(1:20),labels=c(row.names(y.veg),"mu"))
```

```
## Scale for 'y' is already present. Adding another scale for 'y', which will
## replace the existing scale.
```

```
p2 = p2 + theme(axis.text.y=element_text(size=8))
p4 = p4 + scale_y_continuous(breaks=c(1:20),labels=c(row.names(y.veg),"mu"))
```

```
## Scale for 'y' is already present. Adding another scale for 'y', which will
## replace the existing scale.
```

```
p4 = p4 + theme(axis.text.y=element_text(size=8))
# arrange ggplots into grid
grid.arrange(p1, p2, p3, p4, nrow = 2)
```



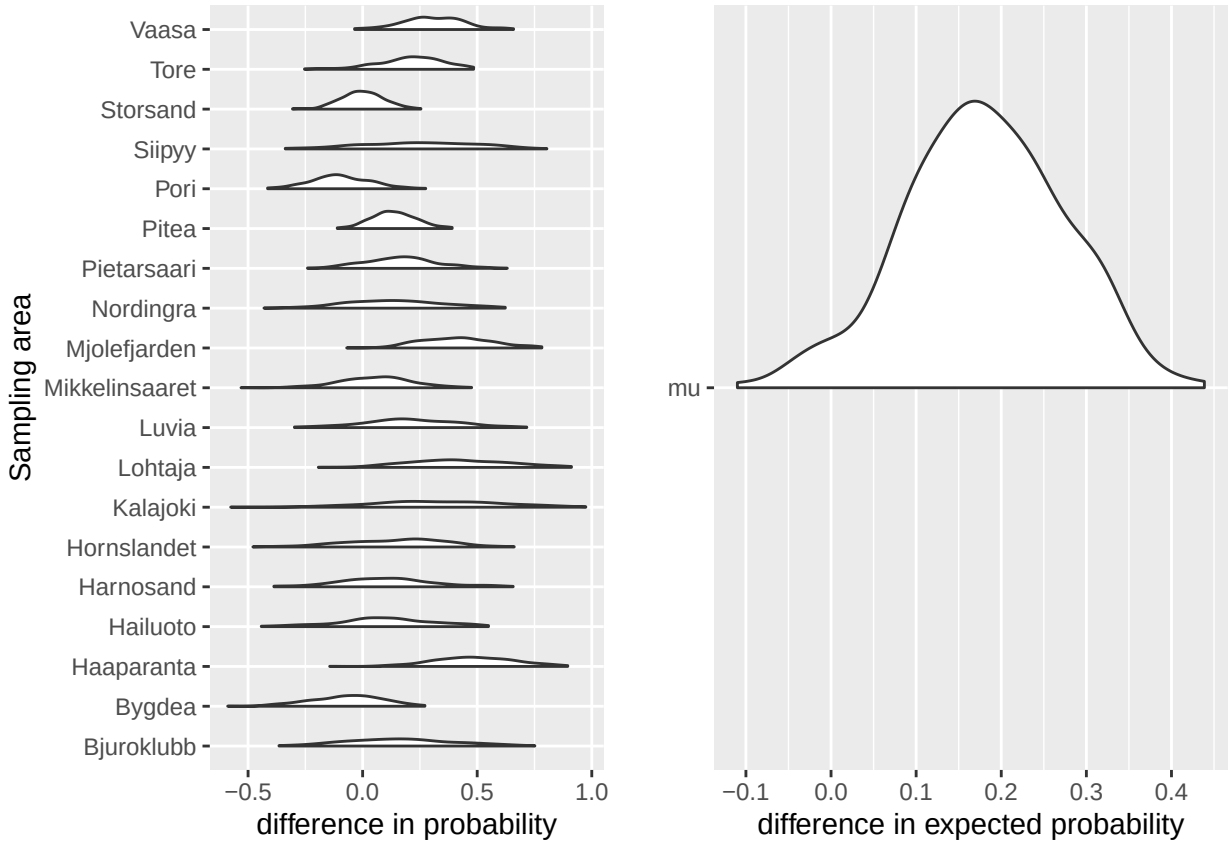
From the above figures we can see that there is considerable variation in probability of white fish larvae presence between sampling areas. However, since the posterior of μ_1 is concentrated in smaller values than the posterior of μ_0 the probability of white fish larvae is smaller in sites with bottom vegetation than in sites with no bottom vegetation. Moreover, by looking at the posterior distribution of s_1 and s_2 we can conclude that the variation in whitefish larvae presence probability across sampling areas is similar in both vegetated

and non-vegetated sites.

2

Let's visualize the posterior distributions of $\Delta\mu = \mu_0 - \mu_1$ and $\phi_i = \theta_{i,0} - \theta_{i,1}$ for each area $i = 1, \ldots, 19$.

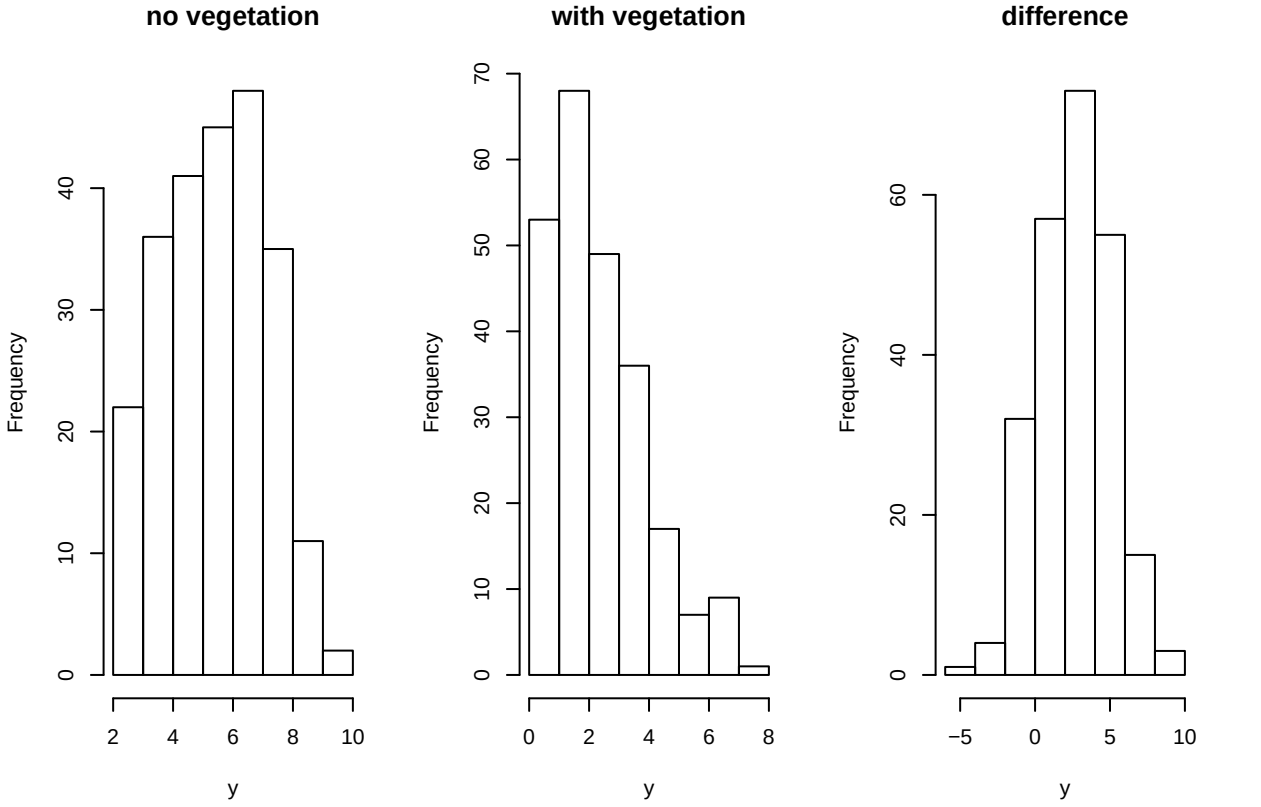
```
# generate samples of phi and Delta-mu
theta.noveg = as.matrix(post.noveg,pars="theta")
theta.veg = as.matrix(post.veg,pars="theta")
phi = theta.noveg - theta.veg
Dmu = as.matrix(post.noveg,pars="mu") - as.matrix(post.veg,pars="mu")
# Set the column names to area names
colnames(phi) <- row.names(y.veg)
# put samples into data frame in order to allow ggplotting
phi.fr = data.frame(name=c( rep(row.names(y.veg),dim(phi)[1])), value=c(t(phi)) )
Dmu.fr = data.frame(name=c( rep("mu",dim(Dmu)[1])), value=c(Dmu) )

p1 = ggplot(phi.fr, aes(x=name,y=value)) + geom_violinhalf() +
   coord_flip() + labs(x="Sampling area", y = "difference in probability")
p2 = ggplot(Dmu.fr, aes(x=name,y=value)) + geom_violinhalf() +
   coord_flip() + labs(x="", y = "difference in expected probability")
grid.arrange(p1,p2,nrow=1)
```



3

Let's calculate the posterior predictive distribution of outcome $\tilde{y}_{19}$ in new sampling with $\tilde{N}_{19} = 10$; that is, the number of new sites in Vaasa area that will have white fish larvae. Since we are predicting into area that is included in our data we we can sample from the posterior of $\tilde{y}_{19}$ directly by using the samples of θ_{19}.

```
y_pred19.noveg = rbinom(dim(theta.noveg)[1],10,theta.noveg[,19])
y_pred19.veg = rbinom(dim(theta.veg)[1],10,theta.veg[,19])
par(mfrow=c(1,3))
hist(y_pred19.noveg, main="no vegetation", xlab="y")
hist(y_pred19.veg, main="with vegetation", xlab="y")
hist(y_pred19.noveg-y_pred19.veg, main="difference", xlab="y")
```

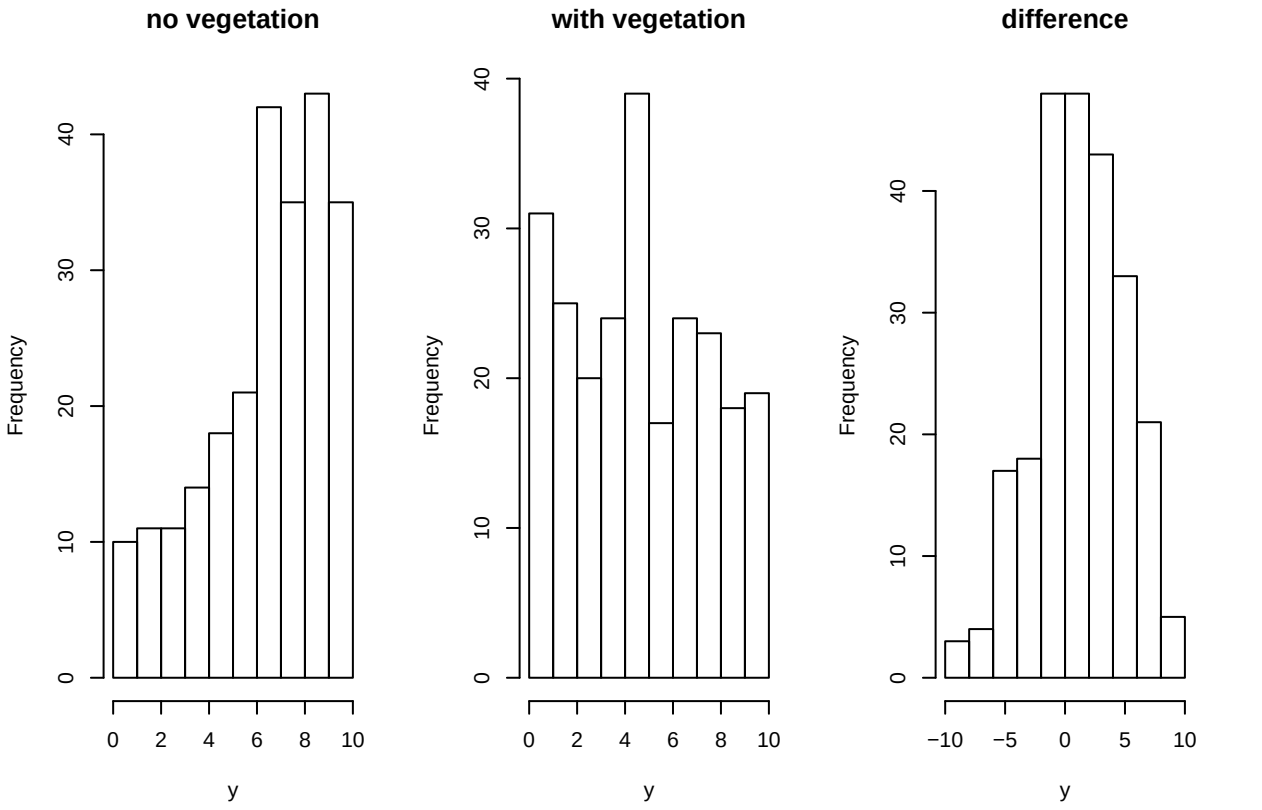


4

In order to sample from the posterior predictive distribution of outcome $\tilde{y}_{20,c}$ of a new sampling with $\tilde{N}_{20} = 10$ in a new sampling area $i = 20$, we need to first sample the $\theta_{20,c}$ parameters for that site. After that the sampling for $\tilde{y}_{20,c}$ goes as above. Hence, the result is

```
mu_s.noveg = as.matrix(post.noveg,pars=c("mu","s"))
mu_s.veg = as.matrix(post.veg,pars=c("mu","s"))
theta20.noveg = rbeta(dim(mu_s.noveg)[1],
                      mu_s.noveg[,"mu"]*mu_s.noveg[,"s"],
                      mu_s.noveg[,"s"]-mu_s.noveg[,"mu"]*mu_s.noveg[,"s"])
theta20.veg = rbeta(dim(mu_s.veg)[1],
                      mu_s.veg[,"mu"]*mu_s.veg[,"s"],
                      mu_s.veg[,"s"]-mu_s.veg[,"mu"]*mu_s.veg[,"s"])
```

```
y_pred20.noveg = rbinom(length(theta20.noveg),10,theta20.noveg)
y_pred20.veg = rbinom(length(theta20.veg),10,theta20.veg)
par(mfrow=c(1,3))
hist(y_pred20.noveg, main="no vegetation", xlab="y")
hist(y_pred20.veg, main="with vegetation", xlab="y")
hist(y_pred20.noveg-y_pred20.veg, main="difference", xlab="y")
```



5

The posterior distribution of the pooled model (exercise 3 of week 2) does not resemble any of the posterior distributions of area-wise $\theta_{i,c}$ above. However, it somewhat resembles the posterior distribution of μ even though it is narrower than that. The reason is that the information from individual areas is stronger for θ of the pooled model than for μ of the hierarchical model since in the hierarchical model some of the information is used for posterior of $\theta_{i,c}$.

The hierarchical model seems more justified since it allows for differences between different areas. Gulf of Bothnia is rather large area so it is very likely that the abundance and due that the probability of presence of white fish varies considerably within it. The hierarchical prior accounts for these variations. In the pooled model single area with very high or low probability of presence can influence the posterior distribution of θ considerably – especially if such an area is sampled more extensively than the other areas. In the hierarchical model, single area does not have such a strong effect and moreover the impact of the differences in area-wise sample sizes (N_i) are not as large.

Grading

Total 20 points Each of the steps provides 4 points from correct answer and 2 points from an answer that is towards the right direction but includes minor mistake (e.g. a bug or typo)

References

Lari Veneranta, Richard Hudd and Jarno Vanhatalo (2013). Reproduction areas of sea-spawning Coregonids reflect the environment in shallow coastal waters. Marine Ecology Progress Series, 477:231-250. http://www.int-res.com/abstracts/meps/v477/p231-250/