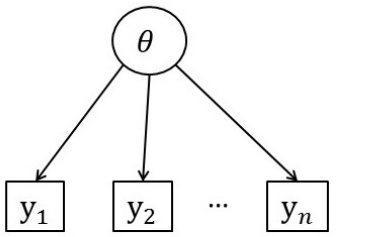
Directed acyclic graphs

Week4-ex1, problem statement

a)-c) 1) Write down the joint probability distribution of all the parameteters and observation variables y in the directed acyclic graphs (DAG) shown in Figure 1, and 2) write down a Stan pseudo code that tells how a model corresponding to that DAG would be written. You can assume that all variables get values in real numbers. An example of a model answer is provided on the right



Joint distribution

$$p(y_1, \ldots y_n, \theta) = p(\theta) \prod_{i=1}^{n} p(y_i|\theta)$$

Stan pseudo-code:

```
data{
   int<lower=0> n;
   real y[n];
}
parameters{
   real theta;
}
model{
   theta ~ p();
   for( i in 1 : n ) {
      y[i] ~ p(theta);
   }
}
```

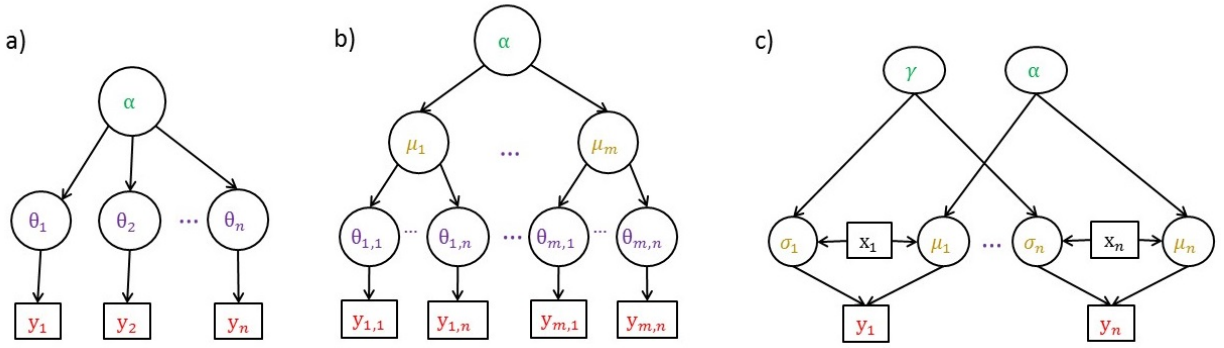


Figure 1: The DAGs for which the joint distribution and pseudo code have to be defined. Note variables denoted by x should be treated as covariates.

d) Draw a Directed acyclic graph (DAG) and write a Stan pseudo code of the following model

$$\begin{aligned}
y_{i,j} &\sim N(\mu_j, \sigma_j^2), i = 1, \ldots, n, j = 1, \ldots, J \\
\mu_j &\sim N(\mu_0, \phi) \\
\mu_0 &\sim N(0, 10^6) \\
\phi &\sim \text{Inv-}\chi^2(\nu_1, s_1^2) \\
\sigma_j^2 &\sim \text{Inv-}\chi^2(\nu_2, s_2^2)
\end{aligned}$$

See the previous problem for an example on the needed accuracy for the pseudo code.

Grading

Total 20 points. a)-c) Two points for correct joint density function and 3 points for correct pseudo-code. d) 3 points for correct DAG and 2 points for correct pseudo-code.

Effect of bottom coverage to larval presence

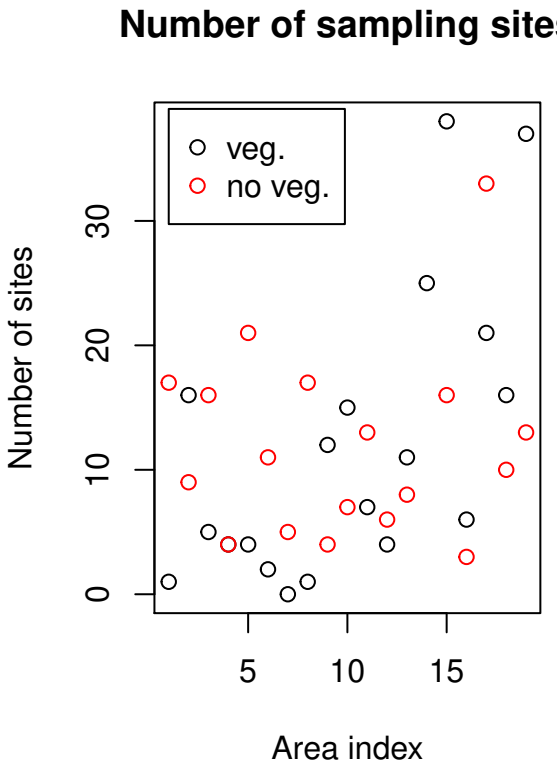
Week4-ex3, solution

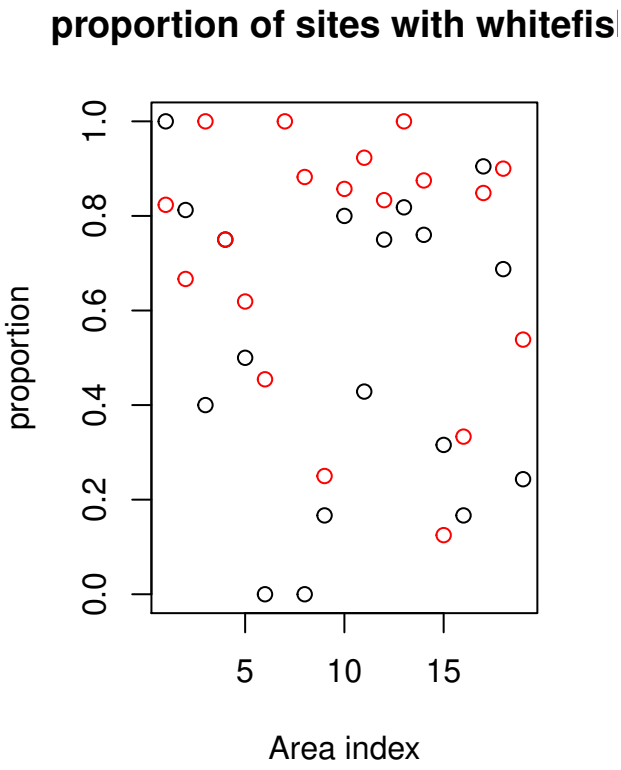
In this exercise, we continue the analysis of the white fish larval areas (week 2, exercise 3). We are again interested in analysing whether or not bottom vegetation affects white fish larvae occurrence probability. However, instead of having a common probability of presence parameter across the Gulf 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?

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/