BDA - Assignment 2

Anonymous

Contents

Exercise 1) Inference for binomial proportion

a)

π is the probability of a monitoring site having detectable algae levels. The prior distribution is modeled by a beta distribution Beta(2,10) as

$$p(\pi) \propto \pi^{\alpha-1}(1-\pi)^{\beta-1} = \pi^1(1-\pi)^9.$$

The posterior density for π is

$$p(\pi|y) = \text{Beta}(\pi|\alpha + y, \beta + n - y) = \text{Beta}(\pi|2 + 44, 10 + 274 - 44) = \text{Beta}(46, 240),$$

based on the derivation of the posterior in p.35 of the BDA3 and our data:

```
library(aaltobda)
data("algae")

# test data
algae_test <- c(0, 1, 1, 0, 0, 0)

n <- length(algae)
n
```

```
## [1] 274
```

```
y <- sum(algae)
y
```

```
## [1] 44
```

The likelihood $p(y|\pi)$ is given by

$$p(y|\pi) = \binom{n}{y}\pi^y(1-\pi)^{n-y} = \binom{274}{44}\pi^{44}(1-\pi)^{230}.$$

Thus, the likelihood is parametrized by Beta(43,240).

b)

According to the observations and the prior knowledge, we can obtain a point estimate $E(\pi|y)$ by taking the posterior mean of π

```
alpha <- 2
beta <- 10
beta_point_est <- function(prior_alpha, prior_beta, data) {
  n <- length(data)
  y <- sum(data)

  return((prior_alpha+y)/(prior_alpha+prior_beta+n))
}
beta_point_est(prior_alpha = alpha, prior_beta = beta, data = algae)
```

```
## [1] 0.1608392
```

We can also obtain the 90% posterior interval

```
beta_interval <- function(prior_alpha, prior_beta, data, prob) {
  n <- length(data)
  y <- sum(data)

  # posterior parameters
  posterior_alpha <- prior_alpha + y
  posterior_beta <- prior_beta + n - y

  # lower and upper intervals
  lower <- (1-prob)/2.0
  upper <- 1-lower

  return (qbeta(c(lower, upper), posterior_alpha, posterior_beta))
}
beta_interval(prior_alpha = 2, prior_beta = 10, data = algae, prob = 0.9)
```

```
## [1] 0.1265607 0.1978177
```

c)

The probability that the proportion of monitoring sites with detectable algae levels π is smaller than $\pi_0 = 0.2$ that is known from historical records is

```
beta_low <- function(prior_alpha, prior_beta, data, pi_0) {
  n <- length(data)
  y <- sum(data)

  # posterior parameters
  posterior_alpha <- prior_alpha + y
  posterior_beta <- prior_beta + n - y

  return(pbeta(pi_0, posterior_alpha, posterior_beta))
}
beta_low(prior_alpha = 2, prior_beta = 10, data = algae, pi_0 = 0.2)
```

```
## [1] 0.9586136
```

d)

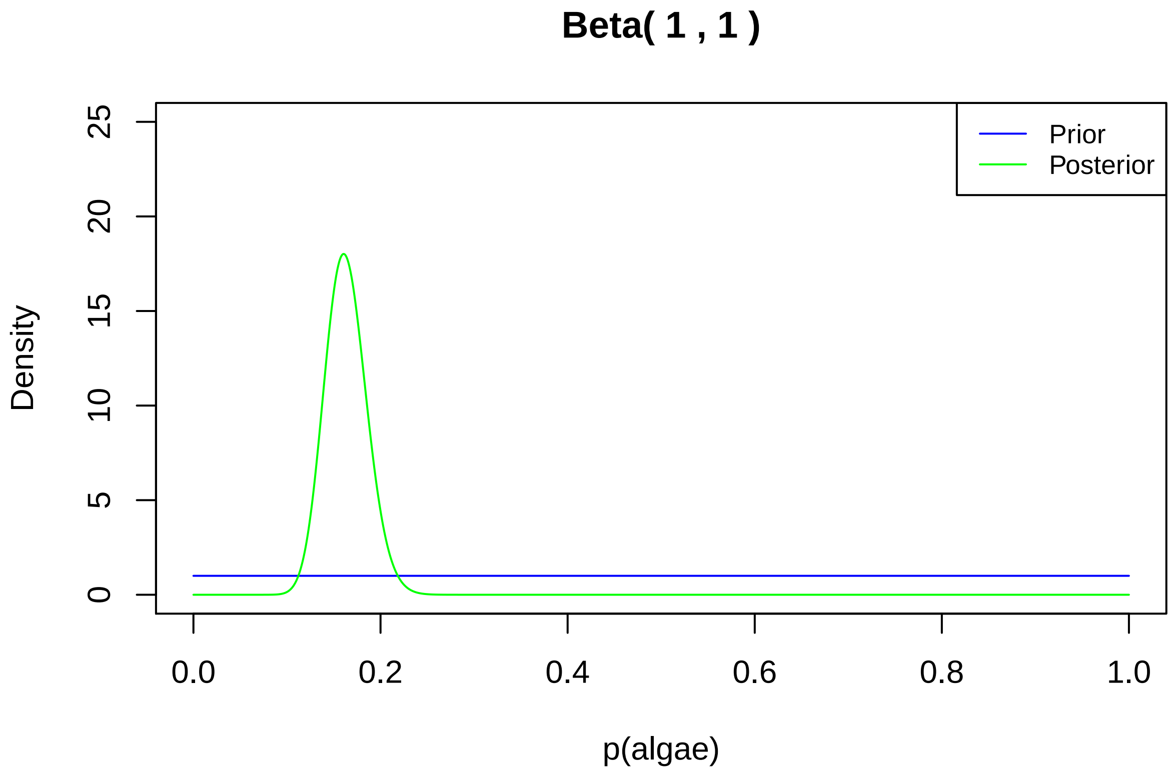
The assumptions needed to use this kind of model with this kind of data is that our data (i.e. `algae`) are independent and identically distributed: all data points stem from the same generative process (same distribution) and the sampled data points have no impact in the future samples.

Furthermore, there must be two events possible, one with probability θ and another with probability $1 - \theta$.

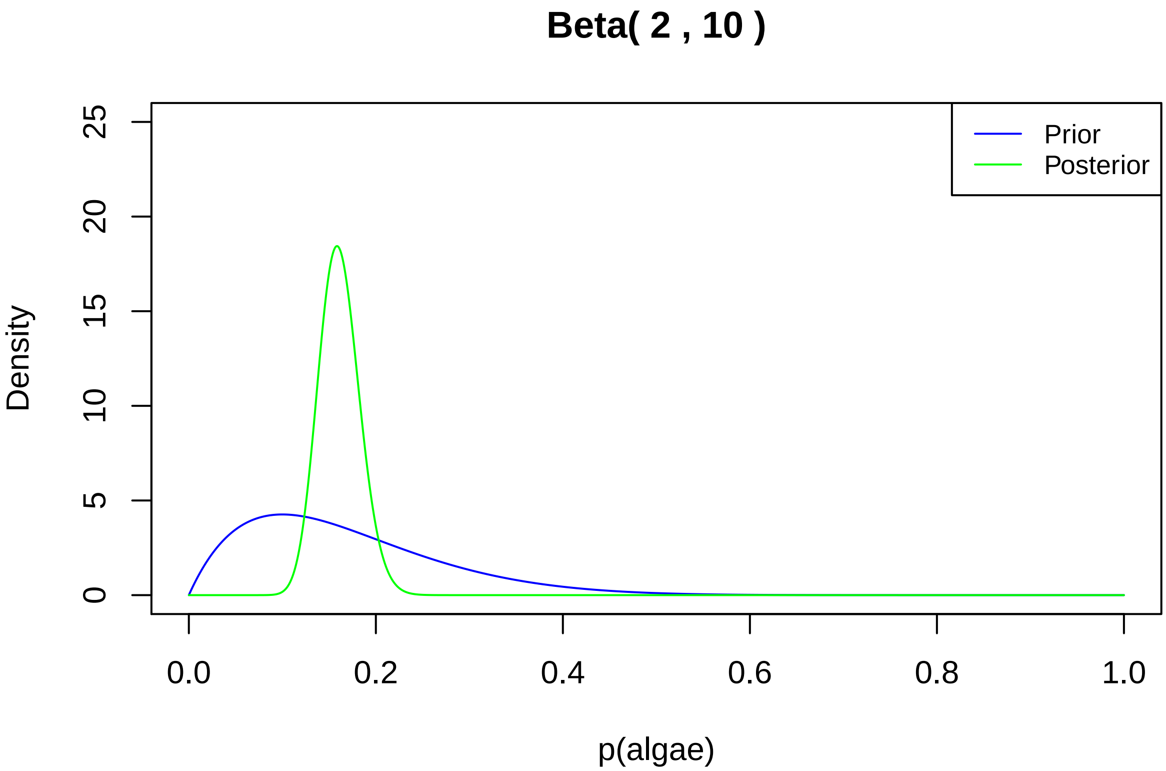
e)

In this case, we will try a uniform Beta(1,1) prior, the previous Beta(2,10) prior, a Beta (2, 30) prior and another Beta(20, 100) prior.

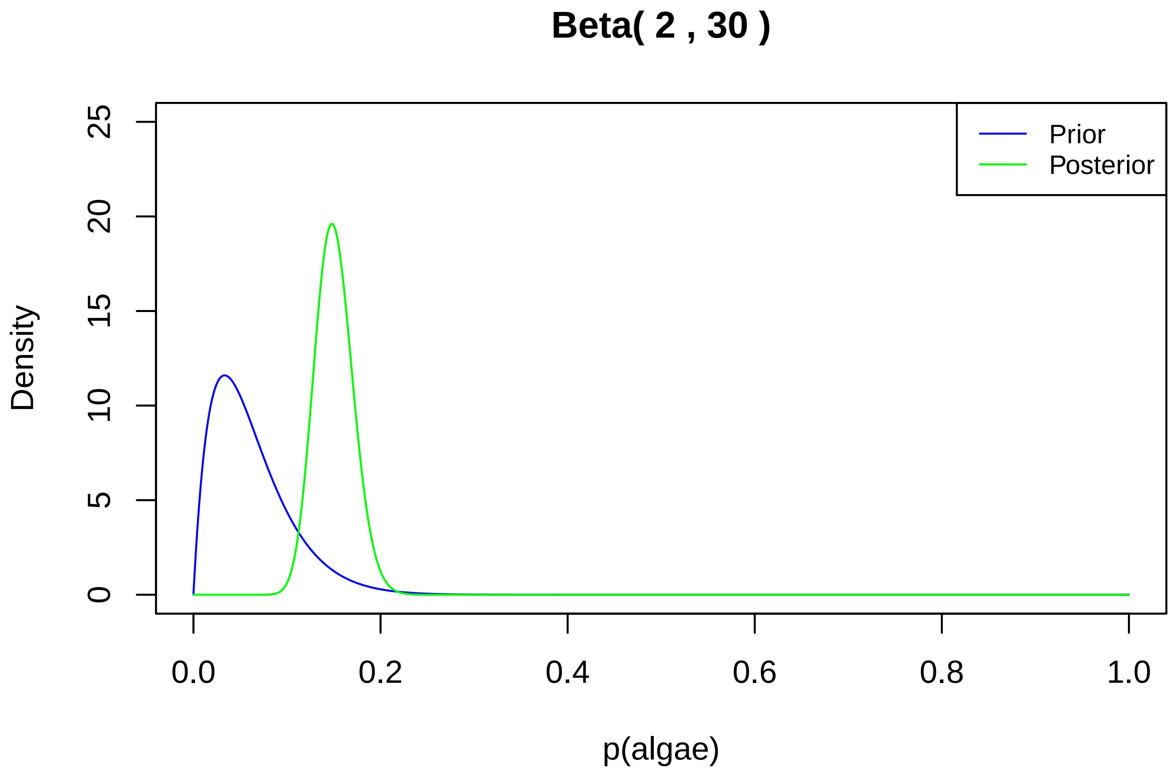
```r
plot_posterior <- function(prior_alpha, prior_beta, data) {
  n <- length(data)
  y <- sum(data)

  # posterior parameters
  posterior_alpha <- prior_alpha + y
  posterior_beta <- prior_beta + n - y

  x <- seq(0, 1, length = 10000)
  density <- dbeta(x, prior_alpha, prior_beta)
  plot(x, density, type = "l", col = "blue", xlab = "p(algae)", ylab = "Density", ylim = c(0, 25), main

  density <- dbeta(x, posterior_alpha, posterior_beta)
  lines(x, density, type = "l", col = "green")
  legend("topright", legend = c("Prior", "Posterior"),
         col = c("blue", "green"), lty = 1, cex = 0.8)
}
plot_posterior(prior_alpha = 1, prior_beta = 1, data = algae)
```



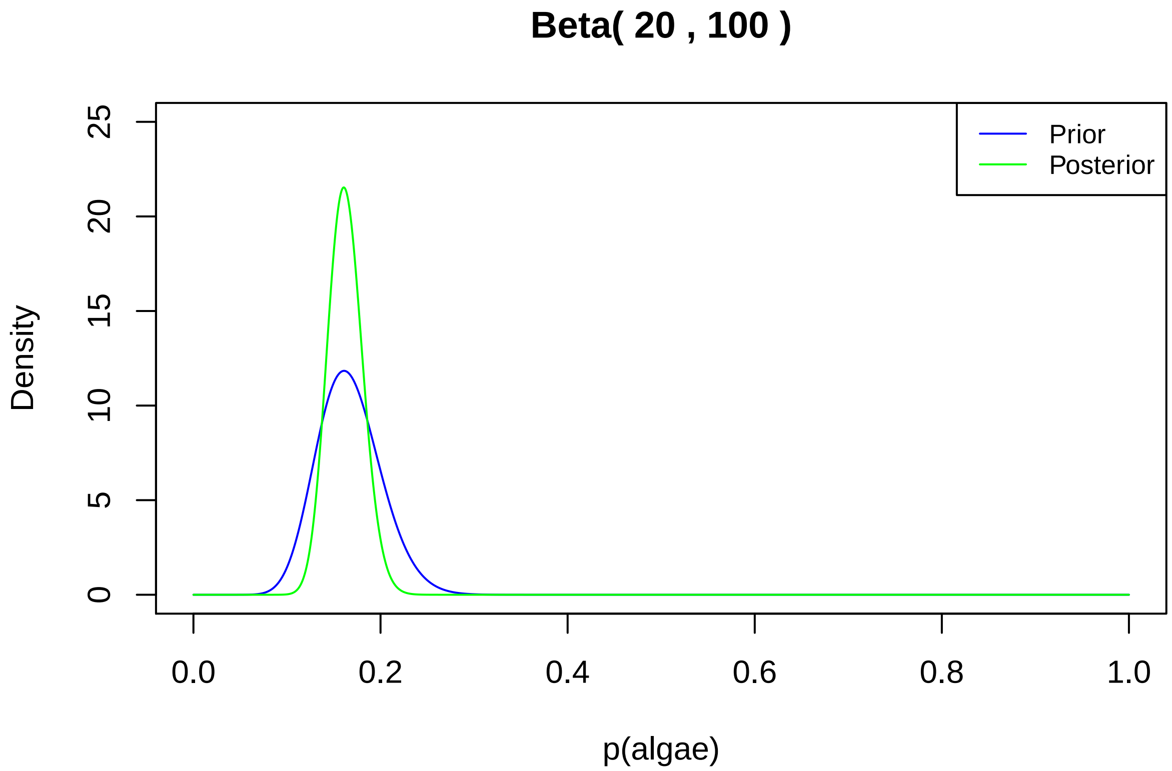
```r
plot_posterior(prior_alpha = 2, prior_beta = 10, data = algae)
```



```
plot_posterior(prior_alpha = 2, prior_beta = 30, data = algae)
```

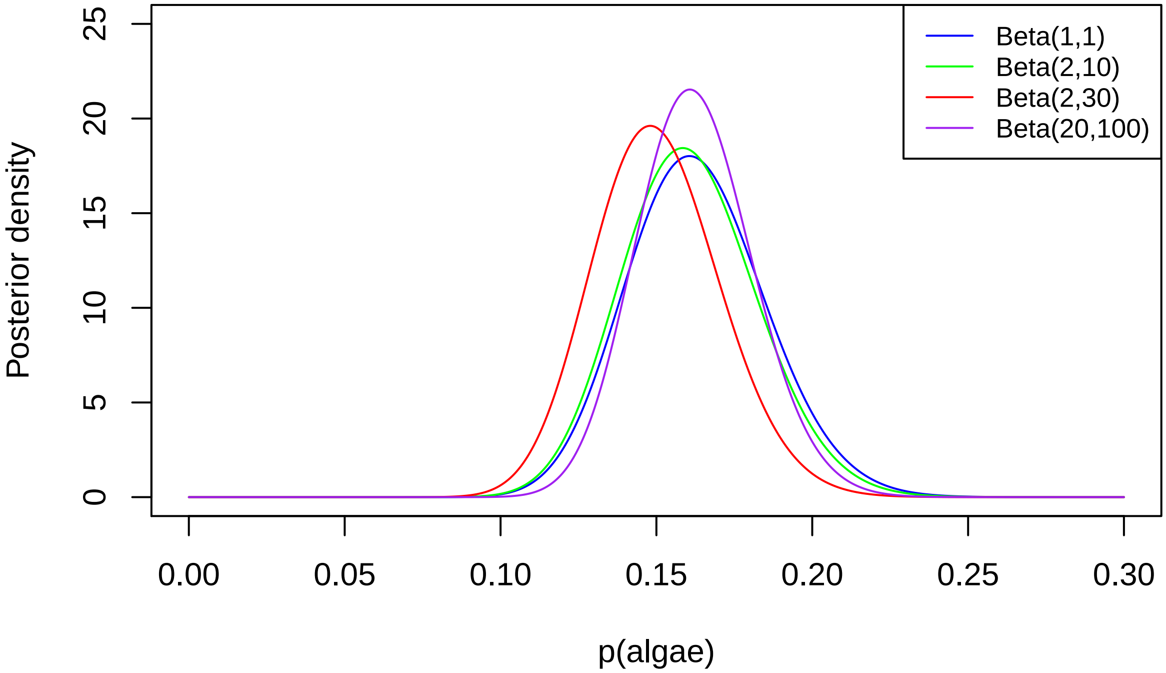


```
plot_posterior(prior_alpha = 20, prior_beta = 100, data = algae)
```



To compare the result of these priors into the different posteriors we can plot all the posteriors together.

```
n <- length(algae)
y <- sum(algae)

prior_alpha = 1
prior_beta = 1
x <- seq(0, 0.3, length = 10000)
density <- dbeta(x, prior_alpha + y, prior_beta + n - y)
plot(x, density, type = "l", col = "blue", xlab = "p(algae)", ylab = "Posterior density", ylim = c(0, 25

prior_alpha = 2
prior_beta = 10
density <- dbeta(x, prior_alpha + y, prior_beta + n - y)
lines(x, density, type = "l", col = "green")

prior_alpha = 2
prior_beta = 30
density <- dbeta(x, prior_alpha + y, prior_beta + n - y)
lines(x, density, type = "l", col = "red")

prior_alpha = 20
prior_beta = 100
density <- dbeta(x, prior_alpha + y, prior_beta + n - y)
lines(x, density, type = "l", col = "purple")

legend("topright", legend = c("Beta(1,1)", "Beta(2,10)", "Beta(2,30)", "Beta(20,100)"),
       col = c("blue", "green", "red", "purple"), lty = 1, cex = 0.8)
```



Overall, we do not see large differences in the observed posteriors. The largest difference in the point estimate is when using Beta(2, 30) as a prior. This is because, as can be seen in the previous plots, this prior is very skewed towards zero and has a very narrow distribution compared to the rest of the priors. We can see that the uniform prior (i.e. Beta(1,1)) and Beta(2,10) show similar posteriors as their priors are the most spread (especially in the case of the uniform prior). The narrowest posterior is obtained with Beta(20,100) as prior, which is expected since the prior has a narrow distribution and the posterior is narrower (see previous plots) by definition. This is because, as we get more information, we become more certain of our prediction.