Logit Regression

Javier Esteban Aragoneses, Mauricio Marcos Fajgenbaun, Danyu Zhang, Daniel Alonso

March 21th, 2021

Introduction

The objective of this research is to obtain the best model to predict concentration of protein in blood utilizing genetic SNP data from 100 individuals.

Model

To know what the best models are, we use a Robust prior, this is a prior of the form:

$$\begin{aligned}\pi_i^R(\beta_0, \beta_i, \sigma) &= \pi(\beta_0, \sigma) \times \pi_i^R(\beta_i|\beta_0, \sigma) \\ &= \sigma^{-1} \times \int_0^\infty \nu_{k_i}(\beta_i|0, \Sigma_i) p_i^R(g) dg\end{aligned}$$

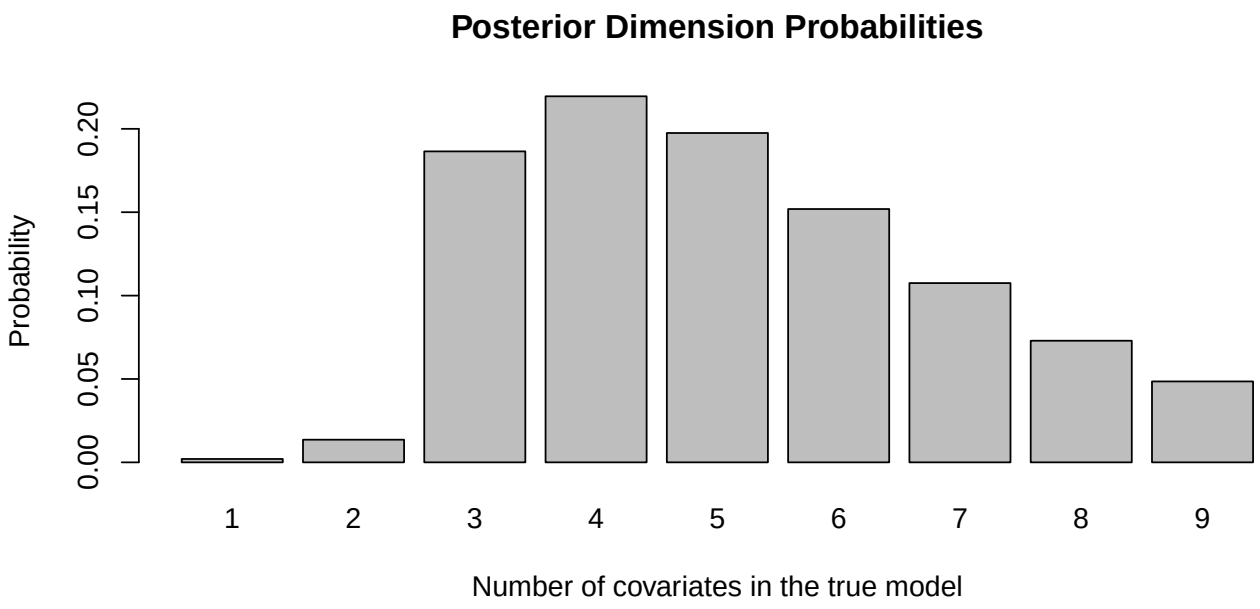
We have chosen this specific prior because its importance is limited. Given a scenario where we have a conflict between the data and the prior, a robust prior gives us significant flexibility when attempting to solve this conflict by attributing more importance to the data.

```
gen <- read.csv("gendata.csv",header = TRUE,sep=";",row.names = 1)
gen$conc <- log(gen$conc)
for (i in 1:(length(names))) {gen[,i] <- as.numeric(gen[,i])}
factor_model <- Bvs(formula= conc ~ ., data=gen, prior.betas = "Robust")
#> Info. . . .
#> Most complex model has 9 covariates
#> From those 1 is fixed and we should select from the remaining 8
#> snp1, snp2, snp3, snp4, snp5, snp6, snp7, snp8
#> The problem has a total of 256 competing models
#> Of these, the  10 most probable (a posteriori) are kept
#> Working on the problem...please wait.
```

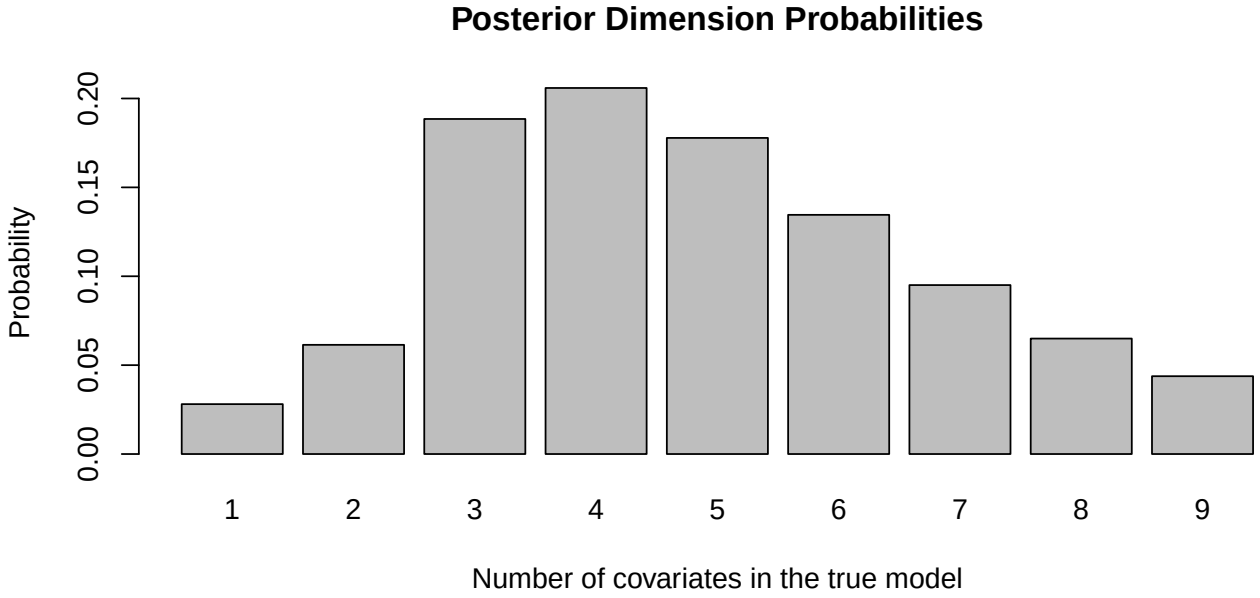
```
gen <- read.csv("gendata.csv",header = TRUE,sep=";",row.names = 1)
for (i in 1:(length(names))) {gen[,i] <- as.factor(gen[,i])}
gen$conc <- log(as.numeric(gen$conc))
numeric_model <- Bvs(formula= conc ~ ., data=gen, prior.betas = "Robust")
#> Info. . . .
#> Most complex model has 9 covariates
#> From those 1 is fixed and we should select from the remaining 8
#> snp1, snp2, snp3, snp4, snp5, snp6, snp7, snp8
#> The problem has a total of 256 competing models
#> Of these, the  10 most probable (a posteriori) are kept
#> Working on the problem...please wait.
```

#From the histogram of factor model, we can observe that #rve that

```
plot(factor_model, option="dimension")
```




```
plot(numeric_model, option="dimension")
```



Models which use predictors of factor type tend to explain better than numeric types models. The reason for this is the fact that the distance between the different values of the genetic profiles are the same. Which means that, for instance, if we use it as quantitative variable, the distance between 1-3 will be larger than 1-1.

As a result, our chosen model is our model which uses factors:

```
# step(lm(formula= conc ~ ., data=gen), k = log(nrow(gen)))
```

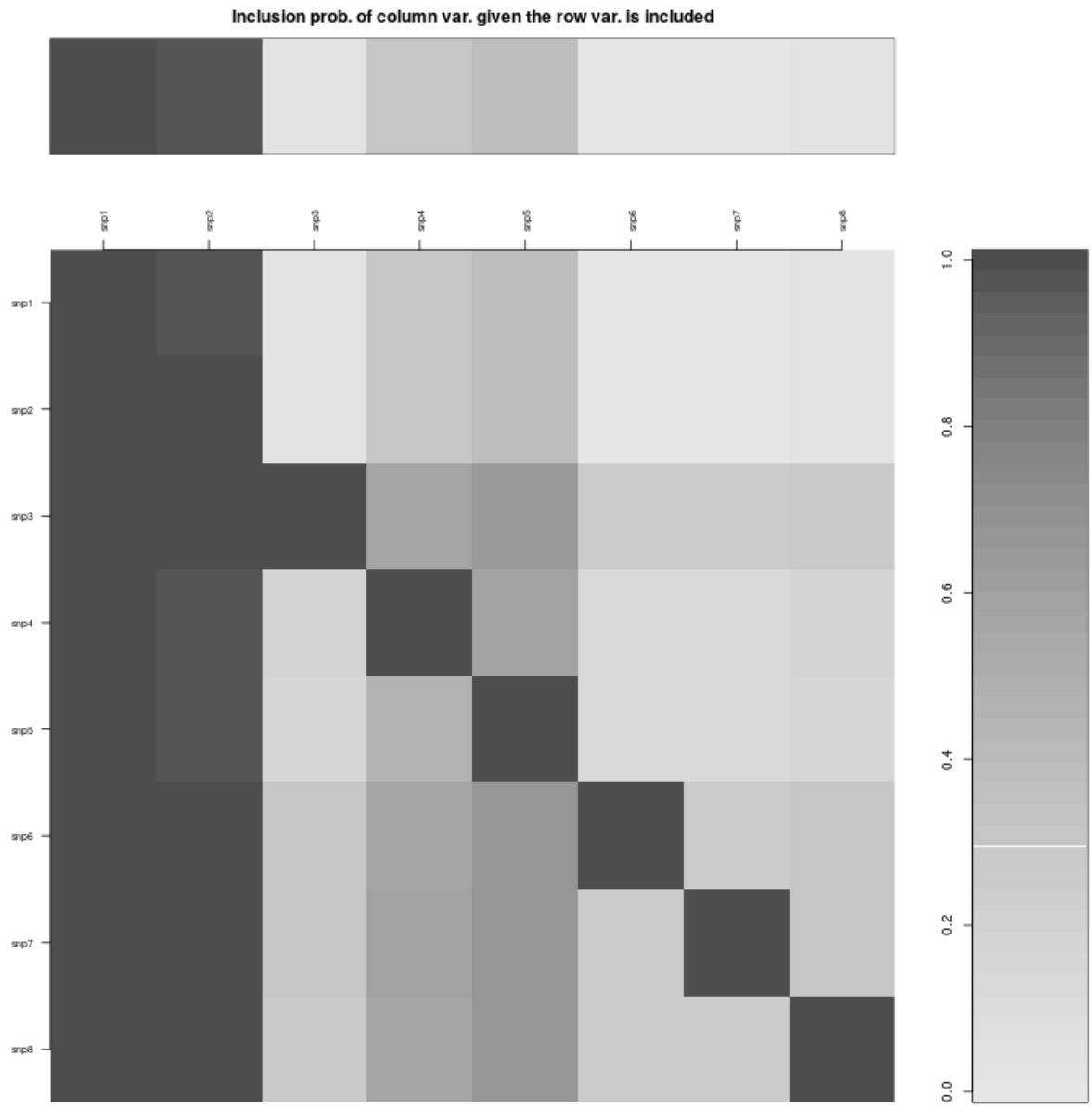


Figure 1: Best models for predictors as factors

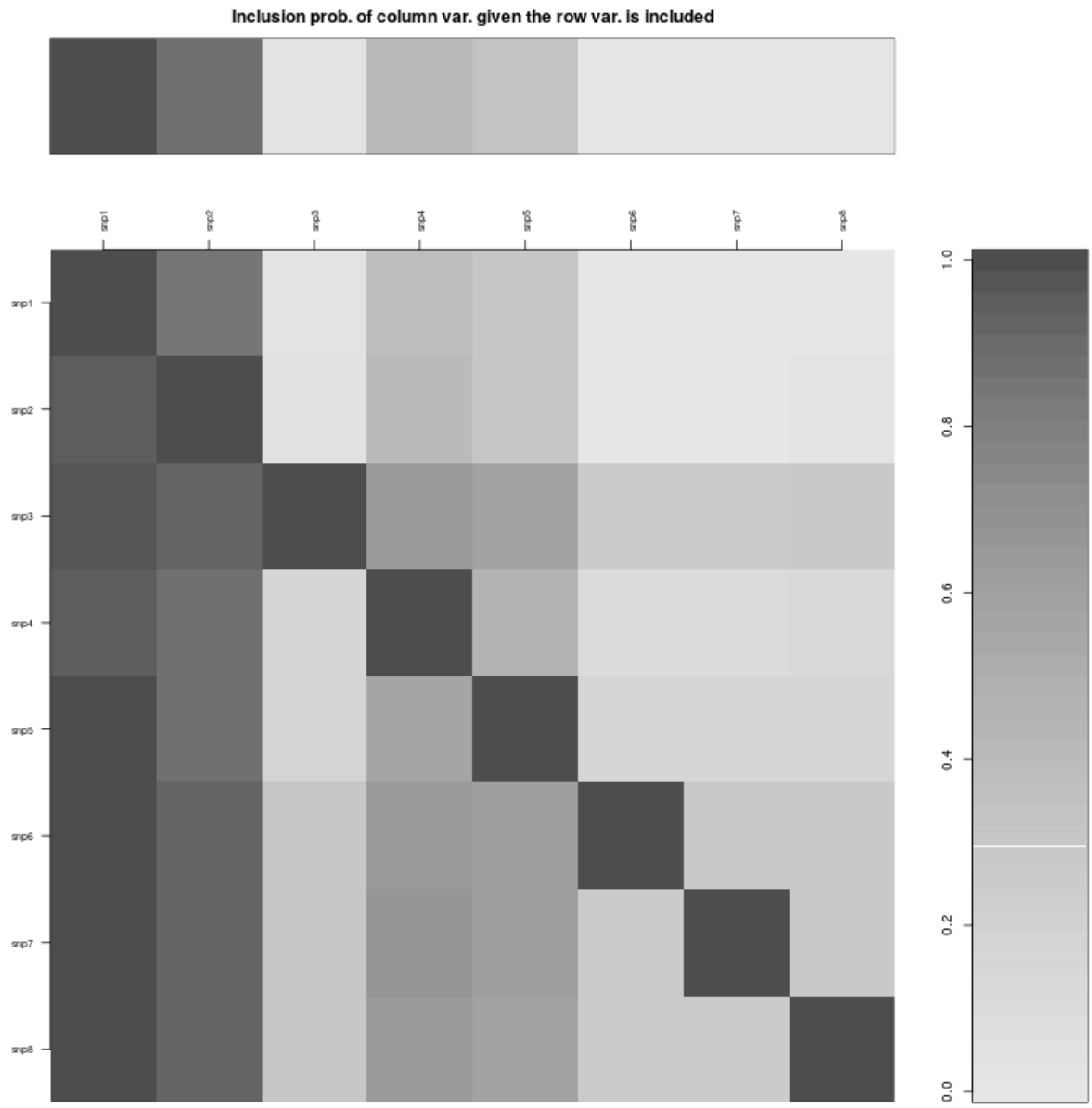


Figure 2: Best models for predictors as continuous values