Bayesian inference 2017
Exercise session 5 (4.–7.12.2017)

1. Let's revisit posterior inference for the normal distribution with both unknown mean and the variance. In the lectures we derived a posterior for this model with the noninformative prior $p(\mu, \sigma^2) \propto (\sigma^2)^{-1}$. Let's now derive the posterior using the general form of the conjugate prior. So our model is now:

$$\begin{aligned} Y_i &\sim N(\mu, \sigma^2) \quad \text{for all } i = 1, \ldots, n, \\ \mu \,|\, \sigma^2 &\sim N(\mu_0, \sigma^2/\kappa_0), \\ \sigma^2 &\sim \text{Inv-}\chi^2(\nu_0, \sigma_0^2). \end{aligned}$$

Now the prior for the parameter (μ, σ^2) is set hierarchically:

$$p(\mu, \sigma^2) = p(\mu|\sigma^2)p(\sigma^2),$$

where the variance has an inverted chi-squared distribution, and mean has a normal distribution given the variance.

(a) Show that this prior distribution is of the form:

$$p(\mu, \sigma^2) \propto (\sigma^2)^{-(\nu_0+3)/2} \exp\left\{-\frac{\nu_0\sigma_0^2 + \kappa_0(\mu_0 - \mu)^2}{2\sigma^2}\right\}.$$

(b) Let's denote this two-dimensional four-parameter distribution (so-called normal inverse chi-squared distribution) as:

$$(\mu, \sigma^2) \sim N\text{-Inv-}\chi^2(\mu_0, \sigma_0^2/\kappa_0, \nu_0, \sigma_0^2).$$

Show that the joint posterior distribution $p(\mu, \sigma^2|\mathbf{y})$ has the same form:

$$(\mu, \sigma^2) \sim N\text{-Inv-}\chi^2(\mu_n, \sigma_n^2/\kappa_n, \nu_n, \sigma_n^2),$$

where

$$\begin{aligned} \mu_n &= \frac{\kappa_0\mu_0 + n\bar{y}}{\kappa_0 + n} \\ \kappa_n &= \kappa_0 + n \\ \nu_n &= \nu_0 + n \\ \nu_n\sigma_n^2 &= \nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{y} - \mu_0)^2. \end{aligned}$$

Here

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

is a sample mean, and

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

is a sample variance. Notice again how the posterior parameters are the convex combinations of the prior parameters and the sample summary statistics computed from the observations.

2. For the rest of the problems of this exercise set we will examine a data set of the baseball statistics of season 1970. This classical example of the shrinkage estimation first appeared in [1]. Here we are going give a fully Bayesian treatment for this problem.

A file `baseball75.csv` contains hit statistics for 18 players. The data set contains the following variables for 18 players:

- `Hits` : number of hits for 45 first at-bats of the season; denoted as $y_1, \dots, y_{18}$.
- `RemainingHits` : number of hits for the rest of the seasons at-bats; denoted as $\tilde{y}_1, \dots, \tilde{y}_{18}$.
- `RemainingAB` : number of at-bats at the rest of the season; denoted as $m_1, \dots, m_{18}$.

The quantity of interest is the batting average, which is the number of hits y_j of the player divided by the number of at-bats n_j of the player:

$$\mathrm{BA}_j = \frac{y_j}{n_j},$$

or rather the true hitting precision θ_j of the player; batting average is maximum likelihood estimate of this quantity.

We are going to predict batting averages of the players for the rest of the season based on the batting averages of the first 45 at-bats. So the column `Hits` is a training set on which we are going to fit our model, and the columns `RemainingHits` and `RemainingAB` are a test set on which we are going to test the fit of our model.

First we will fit two simplest models: the no-pooling model and the complete-pooling model. These are conjugate models which we have already solved, so you can use the analytical results to get the posterior and the posterior predictive distribution, and use `R` only to compute the quantities of interest and draw the figures.

(a) Fit the no-pooling model, where the players are modeled as completely independent:

$$\begin{aligned} Y_j \,|\, \theta_j &\sim \mathrm{Binom}(45, \theta_j) \\ \theta_j &\sim \mathrm{Beta}(\alpha_j, \beta_j) \quad \text{for all } j = 1, \dots, 18. \end{aligned}$$

This means that you can compute the posterior for the true hitting precisions θ_j of each player separately using the beta-binomial model from the first example of this course. You can use for example uniform distribution $\mathrm{Beta}(1,1)$ as prior for each of the players (so that $\alpha_j = \beta_j = 1$ for all $j = 1, \dots, n$).

Draw a picture (cf. Figure 5.4 at page 113 of Bayesian data analysis - book), where on the x-axis are the observed batting averages

$$\mathrm{BA}_j = \frac{y_j}{45},$$

and on the y-axis are posterior medians for the the true hitting precision θ_j (some of the players have same number of hits, so there are less than 18 dots). Draw also 50% equal-tailed credible intervals for the θ_j:s, and a line in 45 degree angle going through the origin into the picture. An example (a similar picture drawn in 2b is in Figure 1).

(b) Fit the complete pooling model, where the true hitting precision is assumed to be same for all of the players:

$$\begin{aligned} Y_j &\sim \mathrm{Binom}(45, \theta) \quad \text{for all } j = 1, \dots, 18 \\ \theta &\sim \mathrm{Beta}(\alpha, \beta). \end{aligned}$$

You can again use for example the uniform distribution Beta$(1, 1)$ as a prior for the common hitting precision θ.

Compute the posterior distribution for the common hitting precision θ, and draw a figure similar to the previous one (example seen in Figure 1). Add also the horizontal line on the posterior median of θ into this and the previous (and the following two) figure.

3. It feels quite a simplification to model the hitting precisions of the players either as being the same, or being completely independent. So let's fit a proper hierarchical model (a.k.a. partial pooling model) in which we model them as being a sample from the same underlying *population distribution*. Now we can estimate also the parameters of this sampling distribution from the observations, so that the amount of shrinkage of the precisions of the single players towards the mean of the players is determined (mostly) by the data, and not by our prior assumptions.

In problem 4 of the exercise set 3 we fit a hierarchical model by a so called *empirical Bayes* procedure, in which we did not set a prior distribution for the parameters of the population distribution, but used point estimates estimated from the data for them. This was little bit of the "double counting" because we first used data to estimate the prior parameters, and then to fit the model. Empirical Bayes is often an useful approximation of the fully Bayesian model, but it does not take into account the uncertainty of the estimation of the parameters of the population distribution because it assumes them to be fixed by using the point estimates.

(a) We assume that the true hitting precisions θ_j of the players are a sample from the same beta distribution Beta(α, β). To give a more intuitive interpretation for the prior distributions, let's reparametrize the prior beta distribution by using the expected value $\phi = \frac{\alpha}{\alpha+\beta}$ and the (pseudo-) sample size $\lambda = \alpha + \beta$ of the distribution. A beta distribution models a probability, so a sensible uninformative prior distribution for its expected value is the uniform distribution Beta$(1, 1)$. The pseudo-sample size must be positive, so we can use for example Pareto$(0.1, 1.5)$-distribution, which is very long-tailed as we remember from the last exercise set, as a prior distribution for it. So the full hierarchical model is now:

$$\begin{aligned} Y_j \,|\, \theta_j &\sim \text{Binom}(45, \theta_j), \\ \theta_j \,|\, \lambda, \phi &\sim \text{Beta}(\lambda\phi, \lambda(1-\phi)), \\ \phi &\sim \text{Beta}(1, 1), \;\; \lambda \sim \text{Pareto}(0.1, 1.5), \;\; \lambda > 0.1. \end{aligned}$$

We assume that parameters ϕ and λ are independent, so that the joint prior distribution of the parameters can be presented as a product of the marginal priors for the parameters:

$$p(\phi, \lambda) = p(\phi)p(\lambda).$$

Write this model using Stan modelling language, and use `rstan` to fit the model. Hint: you can define the prior for the parameter λ as `pareto(0.1,1.5)`. Remember also to constrain λ to being greater than 0.1, and ϕ and θ_j:s into the interval $(0, 1)$.

Draw a figure similar to ones drawn in the last exercises (hint: you can use `quantile`-function to estimate the empirical quantiles from the simulated sample). Interpret the figure: why do the posterior medians lie on the same line?

(b) Let's also try a hierarchical model with a different model structure: Now we model a population distribution of the logit-transform

$$\alpha_j = \text{logit}(\theta_j) = \log \frac{\theta_j}{1-\theta_j}$$

with the normal distribution, so that our full model is:

$$\begin{aligned} Y_j \,|\, \alpha_j &\sim \text{Binom}(45, \text{logit}^{-1}(\alpha_j)), \\ \alpha_j \,|\, \mu, \sigma &\sim N(\mu, \sigma^2), \\ \mu &\sim N(-1, 1),\ p(\sigma) \propto N(0,1),\ \sigma > 0. \end{aligned}$$

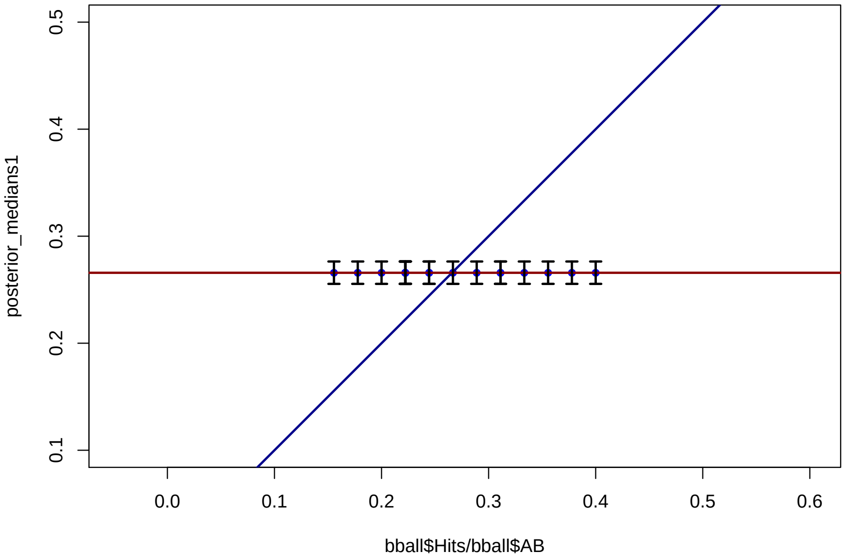
Again we assume that the hyperparameters μ and σ are independent. The prior for the standard deviation σ is a half-normal distribution on the positive real-axis. This works in Stan just by setting `sigma` $\sim$ `normal(0,1)`, and constraining σ to be positive with a constraint `<lower=0>`.

You can get the parameters $\theta_j = \text{logit}^{-1}(\alpha_j) = \frac{1}{1+e^{-\alpha_j}}$ in Stan using a `transformed parameters`-block:

```
transformed parameters {
  real<lower=0, upper=1> theta[n_players] = inv_logit(alpha);
}
```

Draw a picture similar to the previous ones. Which of the hierarchical models shrank the medians of the players more towards the common median?

Kuva 1: Example figure for Problem 2b)



4. Okay, let's see which of our model is best at predicting the batting averages for the rest of the season!

(a) Which of the models do you assume will give the best predictions? How about worst? Why?

(b) The numbers of hits for the rest of season $\tilde{Y}_1, \dots, \tilde{Y}_{18}$ can be modeled as a samples from the binomial distributions having a same success probabilities as the first 45 hits:

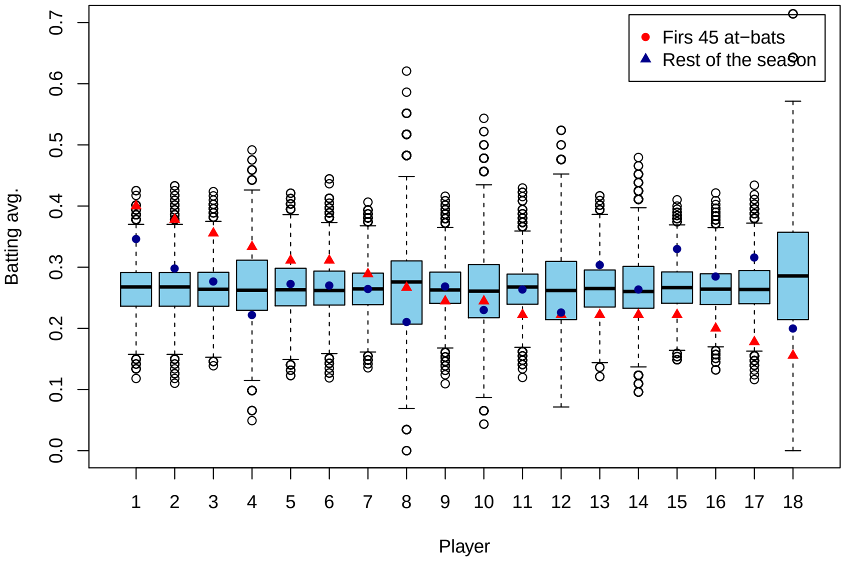
$$\tilde{Y}_j \,|\, \theta_j \sim \text{Binom}(m_j, \theta_j)$$

for each of the players (for the complete-pooling model of course the success probability is same for each of the players: $\theta_j = \theta$ for all $j = 1, \dots, 18$).

Generate samples from the posterior predictive distribution $p(\tilde{\mathbf{y}}|\mathbf{y})$ for each of four models, and based on these draw boxplots of the predicted distribution of the batting averages $BA_j = \tilde{Y}_j / m_j$ for each of the players. Mark also the batting averages for the first 45 at-bats and for the rest of the season for each player into the plot; see Figure 2 for the example plot for the complete pooling model (model of fitted in problem 2b).

For the complete-pooling and no-pooling models you can use just use the function `VGAM::rbetabinom.ab` to simulate from the posterior predictive. For the hierarchical models you can generate a sample from the posterior predictive by generating observations $\tilde{\mathbf{y}}^{(1)}, \dots, \tilde{\mathbf{y}}^{(S)}$ from the sampling distribution of the new observations $p(\tilde{\mathbf{y}}|\boldsymbol{\theta}^{(s)})$ for each $s = 1, \dots, S$, where S is the number of observations simulated from the posterior distribution.

Kuva 2: Example figure for Problem 4b)



(c) Compute the logarithm of the posterior predictive distribution

$$\log p(\tilde{\mathbf{y}}|\mathbf{y}) = \log \left(\prod_{i=1}^{18} p(\tilde{y}_j|\mathbf{y}) \right) = \sum_{j=1}^{18} \log p(\tilde{y}_j|\mathbf{y})$$

(often this is little bit informally called the log likelihood) for the really observed hits $\tilde{y}_1, \dots, \tilde{y}_{18}$ (column `RemainingHits`) for the rest of the season.

For the no-pooling and the complete-pooling models you can compute the values of the posterior predictive distributions directly by using the function[1] `VGAM::dbetabinom.ab`. For the hierarchical models you can approximate the posterior predictive distribution for each of the players as:

$$p(\tilde{y}_j|\mathbf{y}) \approx \frac{1}{S}\sum_{s=1}^{S} p(\tilde{y}_j|\theta_j^{(s)}),$$

where $\theta_j^{(1)}, \ldots, \theta_j^{(S)}$ is a sample from the posterior for the j:th player. Then these can be used to approximate the logarithm of the joint posterior predictive distribution:

$$\log p(\tilde{\mathbf{y}}|\mathbf{y}) = \sum_{j=1}^{18} \log p(\tilde{y}_j|\mathbf{y}) \approx \sum_{j=1}^{18} \log\left(\frac{1}{S}\sum_{s=1}^{S} p(\tilde{y}_j|\theta_j^{(s)})\right).$$

(d) Which of the models was best at predicting (the model which gives the highest probability for the observed data is best at predicting it) the batting averages for the rest of the season? Which was worst? Was the order same that you predicted before actually computing the values? How would you explain the order of the models?

Viitteet

[1] Bradley Efron and Carl Morris. Data analysis using stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319, 1975.

[1] :: is a scope operator in `R`: for example `VGAM::dbetabinom.ab` means the function `dbetabinom.ab` from `VGAM` package.