STK 4021/9021 - Applied Bayesian Analysis and Numerical Methods
Thordis L. Thorarinsdottir
Fall 2014

Final Exam Part 1: Written Project

This is the written project for STK 4021/9021, fall semester 2014. It is made available on the course website on **Sunday, November 23 at 9:00** and candidates must submit their written reports by **Thursday, November 27 at 10:00**. Reports can be submitted to the reception office at the Department of Mathematics in duplicate, or via email to `thordis@nr.no`. Reports may be written in Norwegian or English and should preferable be text processed (e.g. in LaTeX or Word), though handwritten reports will also be accepted. Relevant figures should be included in the report and computer code (such as `R` or `matlab` code) should also be included, e.g. in an appendix. Give your student number on the first page and state any references you use. Candidates are required to work independently.

The exam set contains three problems and comprises 3 pages. STK 9021 students need to solve all three problems, while STK 4021 students need to solve (any) two out of the three problems.

The data sets needed to solve the problems are available for download at `http://www.uio.no/studier/emner/matnat/math/STK4021/h14/eksamen/`.

Problem 1 (**hierarchical modeling).** The files `school1.dat` through `school8.dat` give weekly hours spent on homework for students sampled from eight different schools. We model the data using a hierarchical normal model

$$\begin{aligned} Y_i|\theta_i &\sim \mathcal{N}(\theta_i, \sigma^2), \quad i = 1, \ldots, 8 \quad \text{iid} \\ \theta_i|\mu, \tau^2 &\sim \mathcal{N}(\mu, \tau^2), \quad i = 1, \ldots, 8 \quad \text{iid.} \end{aligned}$$

(a) Using the prior distributions

$$\begin{aligned} \frac{1}{\sigma^2} &\sim \Gamma\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right) \\ \frac{1}{\tau^2} &\sim \Gamma\left(\frac{\eta_0}{2}, \frac{\eta_0\tau_0^2}{2}\right) \\ \mu &\sim \mathcal{N}(\mu_0, \gamma_0^2) \end{aligned}$$

with parameters

$$\mu_0 = 7, \gamma_0^2 = 5, \tau_0^2 = 10, \eta_0 = 2, \sigma_0^2 = 15, \nu_0 = 2,$$

obtain posterior distributions for the parameters $\{\boldsymbol{\theta}, \sigma^2, \mu, \tau^2\}$.

(b) Run a Gibbs sampling algorithm to approximate the posterior distributions in (a). Assess the convergene of the Markov chain, and find the effective sample sizes for $\{\sigma^2, \mu, \tau^2\}$. Run the chain long enough so that the effective sample sizes are all above 1000.

(c) Compare the posterior densities to the prior densities and discuss what was learned from the data.

(d) Calculate the pooled sample variance of the data. Fix σ^2 as this empirical estimate. Use the empirical Bayes method to estimate μ and τ^2. Find the estimated posterior distribution for θ_i, $i = 1, \ldots, 8$, under this framework.

(e) Plot the prior and the posterior density of $\omega = \frac{\sigma^2}{\sigma^2+\tau^2}$ under the framework in (a) as well as the empirical Bayes estimate obtained in (d). Compare these quantities and describe the evidence for between-school variation.

(f) Obtain the posterior probability that θ_7 is smaller than θ_6, as well as the posterior probability that θ_7 is the smallest of all the θ's both under the framework in (a) as well as the empirical Bayes framework. Compare the results.

Problem 2 (imputation). The file `interexp.dat` contains data from an experiment that was interrupted before all data could be gathered. Of interest was the difference in reaction time of experimental subjects when they were given stimulus A versus stimulus B. Each subject is tested under one of the two stimuli on their first day of participation in the study, and is tested under the other stimulus at some later date. Unfortunately, the experiment was interrupted before it was finished, leaving the researchers with 26 subjects with both A and B responses, 15 subjects with only A responses and 17 subjects with only B responses.

We model the data using a bivariate normal sampling model,

$$\mathbf{Y} = \begin{pmatrix} Y_A \\ Y_B \end{pmatrix} \Big| \theta_A, \theta_B, \rho, \sigma_A^2, \sigma_B^2 \sim \mathcal{N}_2 \left(\boldsymbol{\theta} = \begin{pmatrix} \theta_A \\ \theta_B \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_A^2 & \rho\sigma_A\sigma_B \\ \rho\sigma_A\sigma_B & \sigma_B^2 \end{pmatrix} \right),$$

and assume that the lacking responses are missing at random.

(a) Calculate the empirical estimates of $\theta_A, \theta_B, \rho, \sigma_A^2, \sigma_B^2$ from the data. Use **all** the A responses to get $\hat{\theta}_A$ and $\hat{\sigma_B}$, and use **all** the B responses to get $\hat{\theta}_B$ and $\hat{\sigma}_B$. Use only the complete data cases to get $\hat{\rho}$.

(b) For each person i with only an A response, impute a B response as

$$\hat{y}_{i,B} = \mathbb{E}(Y_B | Y_A = y_{i,A}),$$

where the expectation is taken with respect to the bivariate normal distribution with parameters equal to the estimates from (a). For each person i with only a B response, impute an A response in the same manner.

(c) Using Jeffreys' prior for the parameters, implement a Gibbs sampler that approximates the joint distribution of the parameters and the missing data. Assess the convergence of your algorithm. Compare the posterior distributions for the parameters to the estimates in (a). Compute the posterior mean for $\theta_A - \theta_B$ as well as a 95% posterior confidence interval for $\theta_A - \theta_B$.

(d) Using the same parameter prior as in (c), obtain samples from the posterior distribution of the parameters using the imputed values in (b). Compute the posterior mean for $\theta_A - \theta_B$ as well as a 95% posterior confidence interval for $\theta_A - \theta_B$. Compare with the results in (c).

Problem 3 (linear regression). The file `crime.dat` contains crime rates and data on 15 explanatory variables for 47 U.S. states, in which both the crime rates and the explanatory variables have been centered and scaled to have variance 1. A description of the data can be obtained by typing

```
> library(MASS)
> ?UScrime
```

in `R`.

(a) Fit a normal linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

using the g-prior with $g = n$, $\nu_0 = 2$ and $\sigma_0^2 = 1$. Obtain marginal posterior means and 95% confidence intervals for $\boldsymbol{\beta}$, and compare to the least squares estimates.

(b) Describe the relationships between crime and the explanatory variables. Which variables seem strongly predictive of crime rates?

(c) Randomly divide the crime data roughly in half, into a training set $\{\mathbf{y}_{tr}, \mathbf{X}_{tr}\}$ and a test set $\{\mathbf{y}_{te}, \mathbf{X}_{te}\}$. Using only the training set, obtain least squares regression coefficients $\hat{\boldsymbol{\beta}}_{ols}$. Obtain predicted values for the test data by computing $\hat{\mathbf{y}}_{ols} = \mathbf{X}_{te}\hat{\boldsymbol{\beta}}_{ols}$. Plot $\hat{\mathbf{y}}_{ols}$ versus $\mathbf{y}_{te}$ and compute the mean squared prediction error.

(d) Now obtain the posterior mean $\hat{\boldsymbol{\beta}}_{Bayes} = \mathbb{E}(\boldsymbol{\beta}|\mathbf{y}_{tr})$ using the g-prior described above and the training data only. Obtain predictions for the test set $\hat{\mathbf{y}}_{Bayes} = \mathbf{X}_{te}\hat{\boldsymbol{\beta}}_{Bayes}$. Plot the predictions versus the data, compute the mean squared prediction error and compare to the OLS prediction error. Explain the results.

(e) Repeat the procedures in (b) and (c) many times with different randomly generated test and training sets. Plot the mean squared prediction errors in (b) versus those in (c). Compute the average prediction error for both the OLS and the Bayesian methods. Discuss your results.