STAT 207, Spring 2026
Homework # 4
Released May 14. Due June 1 on Canvas

Instructions: You may discuss the homework problems in small groups, but you must write up the final solutions and codes yourself. You can either type your work in LaTeX or write down on papers and scan. Please make sure that all handwriting are visible and please combine all pages into a single PDF file (in the correct order).

For any problems that involve coding, you must provide written answers and also include your codes. You can either include your codes at the end of the homework and label which questions they correspond to, or include as part of the answer (e.g., in the R Markdown style). You will receive no credit if you submit only codes or only written answers.

For all questions, you are required to implement the sampler yourself by hand without using off-the-shelf packages (e.g., stan, INLA, lme4, nlme, etc.). But you are encouraged to compare your results with those returned by these packages.

1. A researcher records temperature at n locations over T days. Let y_{it} denote the temperature for the ith location on day t. For each location, let $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})$ denote a p-dimensional covariate that is invariant over time. Consider the following regression model

$$y_{it} = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_{it} \quad i = 1, ..., n; t = 1, ..., T$$

Assume heteroscedastic errors, where the variance depends on the location:

$$\epsilon_{it} \overset{ind}{\sim} N(0, \sigma_i^2), \quad i = 1, ..., n; t = 1, ..., T$$

Consider the following standard Bayesian Ridge priors for the coefficients and conjugate priors for the variances:

$$\begin{aligned} \beta_j \mid \tau^2 &\overset{iid}{\sim} N(0, \tau^2), \quad j = 1, ..., p \\ \sigma_i^2 &\overset{iid}{\sim} IG(a_\sigma, b_\sigma), \quad i = 1, ..., n \\ \tau^2 &\sim IG(a_\tau, b_\tau) \end{aligned}$$

where $IG(a, b)$ is the Inverse-Gamma distribution with shape a and scale b.

   (a) (10 pts) Up to a proportional constant, write down the likelihood $p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\sigma}^2)$ using (i) summations involving y_{it}, x_{ij}, and β_j; (ii) matrix representation involving $\boldsymbol{y}$, a design matrix $\boldsymbol{X}$, and a covariance matrix $\boldsymbol{\Sigma}$. Specify $\boldsymbol{y}$, $\boldsymbol{X}$, and $\boldsymbol{\Sigma}$ clearly.

   (b) (10 pts) Write down the posterior joint distribution of all the unknown parameters, up to a proportional constant.

   (c) (15 pts) Obtain and identify the posterior full conditional distributions for $\boldsymbol{\beta}$, σ_i^2, and τ^2.

2. Under the same setting of problem 1. Now consider the following regression model

$$y_{it} = \boldsymbol{x}_i^T \boldsymbol{\beta} + \epsilon_{it} \quad i = 1, ..., n; t = 1, ..., T$$

Assume the independent mixture of double exponential priors for β_j's, i.e.,

$$p(\boldsymbol{\beta} \mid \pi, \tau_0, \tau_1) = \prod_{j=1}^{p} \Big(\pi f(\beta_j \mid \tau_1) + (1 - \pi) f(\beta_j \mid \tau_0) \Big)$$

where $f(\beta_j \mid \tau) = \frac{\tau}{2} e^{-\tau|\beta_j|}$ is the double exponential distribution, and τ_0 and τ_1 are known constants with $\tau_0 > \tau_1 > 0$. Consider the following priors

$$\begin{aligned} \pi &\sim Beta(a, b), \\ \epsilon_{it} &\overset{iid}{\sim} N(0, \sigma^2), \quad i = 1, ..., n; t = 1, ..., T \\ p(\sigma^2) &\propto 1/\sigma^2. \end{aligned}$$

(a) (10 pts) Up to a proportional constant, write down the likelihood $p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\beta}, \sigma^2)$ using (i) summations involving y_{it}, x_{ij}, and β_j; (ii) matrix representation involving $\boldsymbol{y}$ and some design matrix $\tilde{\boldsymbol{X}}$. For (ii) please specify the design matrix and response vector clearly.

(b) (10 pts) A double exponential density can be written as scale mixture of normals due to the following identity

$$\frac{\tau}{2} e^{-\tau|x|} = \int_0^\infty N(x \mid 0, s^2) Exp(s^2 \mid \tau^2/2) ds^2$$

for any $\tau > 0$, where $N(x \mid 0, s^2)$ is the normal density with mean 0 and variance s^2, and $Exp(s^2 \mid \tau^2/2)$ is the exponential density with rate parameter $\tau^2/2$.

Use the above representation to introduce latent variables, for each β_j, that provide a hierarchical formulation of each one of the two components of the prior for β_j. Write down explicitly the resulting hierarchical prior for β_j.

(c) (10 pts) Write down the posterior joint distribution of all the unknown parameters, up to a proportional constant.

(d) (20 pts) Obtain and identify the posterior full conditional distributions for all the parameters of the hierarchical model. Note that you may need the following distribution: x follows a generalized inverse Gaussian distribution $InvGau(a, b, m)$ with $a > 0, b > 0, m \in \mathbb{R}$ if the probability density function is

$$f(x) \propto x^{m-1} \exp\Big(- \frac{ax + b/x}{2} \Big), \quad x > 0.$$

Please specify the name and parameter of any standard distributions.

3. Consider the normal mean model with observations y_i, for $i = 1, ..., n$, from the following model

$$y_i | \theta_i, \sigma^2 \sim N(\theta_i, \sigma^2)$$

with prior $p(\sigma^2) \propto 1/\sigma^2$.

Suppose we know that the observations come from two distinct groups: one group of θ_i's are very close to 0, and the other group has more diffuse distribution. Suppose we assume the following model

$$\theta_i \sim \begin{cases} N(0, \tau_0^2) & \text{with probability } 1-p \\ N(0, \tau_1^2) & \text{with probability } p \end{cases}$$

with known parameters τ_0, τ_1, and we assume a Beta prior on p,

$$p \sim Beta(\alpha, \beta)$$

(a) Using missing data formulation to introduce latent variables indicating which group each observation is from. Write down the hierarchical formulation of the prior on θ using the missing indicator.

(b) Write down the distribution of θ, σ^2, p, and the latent variables given the observed data y, up to a constant.

(c) Identify the full conditionals of all unknown quantities. Please specify the name and parameter of any standard distributions.

4. Simulate data from the following model with $n = 20$ schools, each with $m = 100$ students. For each student, an outcome y_{ij} is observed with a set of $p = 20$ covariates. First, specify some values of $\boldsymbol{\mu}$, $\boldsymbol{\tau}$, and σ^2, and simulate the dataset with

$$\begin{aligned} y_{ij} &= \beta_{0i} + \sum_{k=1}^{p} \beta_{ki} x_{ijk} + \epsilon_{ij}, \quad i = 1, ..., n; j = 1, ..., m \\ \beta_{0i} &\sim N(\mu_0, \tau_0^2), \quad i = 1, ..., n \\ \beta_{ki} &\sim N(\mu_k, \tau_k^2), \quad i = 1, ..., n; k = 1, ..., p \\ \epsilon_{ij} &\sim N(0, \sigma^2) \end{aligned}$$

Fit the following four models: (1) a linear regression with common β_0 and $\{\beta_k\}_{k=1:p}$ for all n groups, (2) a spike-and-slab regression with common β_0 and β_k for all n groups, but assume β_k follows a spike-and-slab mixture model, (3) a random intercept model, and (4) a random slope model. For each model,

(a) Specify the full model with all prior specifications.

(b) Derive and write down the posterior full conditionals.

(c) Fit the model and summarize your posterior distributions of the regression coefficients.

(d) Compare your estimated regression coefficients for each group with the true values. Since there are $n \times (p+1)$ regression coefficients. Please carefully design your visualization or numerical summary to avoid plotting too many figures.

(e) Compare the four models and summarize what you find and whether your findings are as expected from your data generating process.

.

5. In this and next question, we will use a dataset on the box office gross of broadway shows. The dataset can be obtained and processed with the following commands in R

```
library(tidytuesdayR)
library(lubridate)
raw <- tidytuesdayR::tt_load(2020, week = 18)$grosses
raw$year = lubridate::year(raw$week_ending)
data <- subset(raw, year >= 2000)
tab <- table(data$show)
sub <- names(tab)[which(tab > 370)]
data <- subset(data, show %in% sub)
# Change the outcome to million dollar scales
broadway <- data.frame(gross = data$weekly_gross / 1e6,
             show = data$show,
             week_ending = data$week_ending,
             id.show = match(data$show, sub),
             id.year = data$year - 1999,
             id.week = data$week_number)
head(broadway)

# simple scatter plot
library(ggplot2)
ggplot(broadway) + aes(x = week_ending, y = gross) +
        geom_point(size = .5) + facet_wrap(~show)
```

Using the `broadway` dataset, fit a linear mixed effect model in the following form

$$y_{ijk} = \alpha_i + \beta_j + \epsilon_{ijk}$$

where y_{ijk} is the weekly gross for show i at year j and week k, with the following priors

$$\alpha_i \sim N(\mu_\alpha, \tau_\alpha)$$
$$\beta_j \sim N(\mu_\beta, \tau_\beta)$$

Make appropriate choices to specifying any prior parameters or hyperpriors. State and discuss your choice of priors and visualize a posterior summary of the estimated parameters and fitted values. Please do not include big tables (> 20 cells).

6. Using the `broadway` dataset from the previous question, fit an alternative linear mixed effect model in the following form

$$y_{ijk} = \alpha_i + \beta_k + \gamma_i x_{ijk} + \epsilon_{ijk}$$

where y_{ijk} is the weekly gross for show i at year j and week k and $x_{ijk} = j$, i.e., x_{ijk} is the year index from 1 to 21. *Notice that in this question, we assume the outcome is linear in year, instead of treating year as an indicator as in the previous question.* Consider the following priors

$$\alpha_i \sim N(\mu_\alpha, \tau_\alpha)$$
$$\beta_k \sim N(\mu_\beta, \tau_\beta)$$
$$\gamma_i \sim N(\mu_\gamma, \tau_\gamma)$$

Make appropriate choices to specifying any prior parameters or hyperpriors. State and discuss your choice of priors and visualize a posterior summary of the estimated parameters and fitted values. Please do not include big tables (> 20 cells).