Homework 8

STT 465, Bayesian Statistical Methods

Lowell Monis

November 21, 2025

Question 1

The file "azdiabetes.txt" contains data on (diabetes) health-related variables of 532 women living near Phoenix, Arizona. Download the file onto your computer, and import the data into R using:

```
diabetes.data <- read.table("data/azdiabetes.txt", header = TRUE)
y <- diabetes.data[, 2]
X <- cbind(rep(1, length(y)), as.matrix(diabetes.data[, -c(2, 8)]))
```

Then, `y` is the glucose level. And `X` includes the intercept (the first column) and six variables. These six variables (from second to last column) are: number of pregnancies, blood pressure, skin fold thickness, body mass index, diabetes pedigree and age.

(a) 30 points

Fit a linear regression model with glucose level as the response variable and other six variables as the predictor variables (plus the intercept). Using the invariant g-prior with parameter values $g = n = 532$, $a = b = 1$, obtain Monte Carlo approximations to the posterior mean and 95% credible interval for each regression coefficient β_j $(j = 1, 2, \ldots, 7)$.

Answer

First, we set up the parameters and sample storage as defined by the question. We also load the `mvtnorm` package.

```
library(mvtnorm)

# dimensions and invariant g-prior
n <- length(y) -> g
p <- ncol(X)
beta_0 <- rep(0,p) # 7x1 zero vector
Sigma_0 <- diag(p) # 7x7 identity matrix
a <- 1 -> b

S <- 5000 # MCMC samples

# storage
beta_sam <- matrix(0, nrow=S, ncol=p)
sigma2_sam <- numeric(S)
```

We consider the normal linear regression model to be:

$$Y = X\beta + \epsilon$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$.

Further, Y is the $n \times 1$ response vector (glucose level), X is the $n \times p$ data matrix containing the predictors, β is the $p \times 1$ coefficient vector, and σ^2 is the error variance.

The question asks me to use the invariant g-prior with $g = n = 532, a = b = 1$. The invariant g-prior as formulated by Zellner in 1986 is a special case where $\beta_0 = 0$ and $\Sigma_0 = \kappa(X^T X)^{-1}$, with $\kappa = g\sigma^2$.

Under the invariant g-prior, the joint posterior distribution $p(\beta, \sigma^2 \mid y, X)$ can be sampled directly without a Gibbs sampler because $p(\sigma^2 \mid y, X)$ is an Inverse-Gamma distribution.

The sampling process involves two steps:

1. Sample σ^2 from the marginal posterior distribution $p(\sigma^2 \mid y, X) \sim \text{Inverse} - \text{Gamma}(\tilde{a}, \tilde{b})$ where $\tilde{a} = a + \frac{n}{2}$ and $\tilde{b} = b + \frac{1}{2}y^T y - \frac{g}{2(g+1)} y^T X (X^T X)^{-1} X^T y$.
2. Sample β from the conditional posterior distribution $p(\beta \mid y, X, \sigma^2) = \mathcal{N}(\beta_n, \Sigma_n)$ where $\beta_n = \frac{g}{g+1}(X^T X)^{-1} X^T y$ and $\Sigma_n = \frac{g}{g+1}\sigma^2 (X^T X)^{-1}$.

I can now set up the specific parameters and the linear algebra needed. Note that the first column is to simulate a constant, with all values set to 1.

```
# constant quantities
XtX_inv <- solve(t(X) %*% X)
yTy <- sum(y^2)
XtX_inv_Xt_y <- XtX_inv %*% t(X) %*% y

# parameters
beta_n <- (g / (g + 1)) * XtX_inv_Xt_y
Sig_constant <- (g / (g + 1)) * XtX_inv
a_tilde <- a+(n/2)
b_tilde <- b + yTy / 2 -
  (g / (2 * (g + 1))) * (t(y) %*% X %*% XtX_inv %*% t(X) %*% y)
```

We can now use composition sampling to generate 5000 independent Monte Carlo samples from $p(\beta, \sigma^2 \mid y, X)$.

```
set.seed(465)
for (i in 1:S) {
  sigma2_sam[i] <- 1 / rgamma(1, shape = a_tilde, rate = b_tilde)
  Sigma_n <- Sig_constant * sigma2_sam[i]
  beta_sam[i, ] <- rmvnorm(1, mean = beta_n, sigma = Sigma_n)
}
```

The Monte Carlo samples are used to approximate the posterior mean the 95% credible intervals for each regression coefficient β_j. I am also adding a flag that checks whether we are 95% confident that the coefficient is non-zero. A variable is only considered strongly predictive if its 95% credible interval does not include zero, i.e., the coefficient is non-zero by a fair chance and contributes the variable greatly to the final model. Another way to put this is that there is strong evidence that there is evidence of a difference between the two groups, when zero is not included in the credible interval.

```r
posterior_means <- colMeans(beta_sam)
credible_intervals <- t(apply(beta_sam, 2, quantile, probs = c(0.025, 0.975)))

var_names <- colnames(X)
results <- data.frame(
  Variable = var_names,
  Posterior_Mean = round(posterior_means, 4),
  CI_Lower = round(credible_intervals[, 1], 4),
  CI_Upper = round(credible_intervals[, 2], 4),
  CI_Excludes_Zero = (credible_intervals[, 1] > 0)|(credible_intervals[, 2] < 0)
)
results
```

```
##   Variable Posterior_Mean CI_Lower CI_Upper CI_Excludes_Zero
## 1                 52.2568  34.9343  69.6241             TRUE
## 2    npreg        -0.6576  -1.6333   0.3011            FALSE
## 3       bp         0.2073  -0.0175   0.4351            FALSE
## 4     skin         0.1945  -0.1257   0.5053            FALSE
## 5      bmi         0.6384   0.1559   1.1309             TRUE
## 6      ped        10.4500   3.1849  17.8224             TRUE
## 7      age         0.7626   0.4535   1.0751             TRUE
```

Upon running our model, it looks like the intercept, `bmi`, `ped`, and `age` are strongly predictive of glucose levels. This indicates that with all variables assumed zero, there is 95% possibility that the true value of the glucose level is within 34.93 and 69.62. The body mass index, diabetes pedigree, and age are the strongest predictors of glucose levels. Higher these values, the higher the predicted glucose level. All variables as stated above are positively related to the glucose level.

(b) 30 points

Perform the model selection and averaging (no need to add interaction or quadratic terms) using the invariant g-prior with parameter values $g = n = 532$, $a_z = b_z = 1$. Use MCMC approximations to obtain $P(z_j = 1 \mid y, X)$ $(j = 1, 2, \ldots, 7)$ as well as the posterior mean and 95% credible interval for each regression coefficient β_j $(j = 1, 2, \ldots, 7)$. Compare to the results in (a).

Answer

The objective is to perform Bayesian model selection and averaging using the Gibbs sampler to account for uncertainty in which of the seven predictors (intercept + six variables) truly belong in the model. We will simultaneously estimate the posterior probability of each model and the model-averaged posterior mean/coefficient $\hat{\beta}_{\text{MA}}$.

We use the same normal linear regression likelihood as in (1a).

$$Y \mid X, \beta, \sigma^2 \sim \mathcal{N}(X\beta, \sigma^2 I_n)$$

For Bayesian model selection and averaging, we introduce the model indicator $\mathbf{z} = (z_1, \ldots, z_p)$, where $z_j = 1$ if predictor j is included, and $z_j = 0$ if it isn't.

We use the invariant g-prior conditional on a model $\digamma$: $p(\beta, \sigma^2 \mid \mathbf{z}) = p(\beta \mid \sigma^2, \mathbf{z}) \cdot p(\sigma^2 \mid \mathbf{z})$.

The priors are as follows:

$$\sigma^2 \sim \text{Inverse} - \text{Gamma}(a_z, b_z) = \text{Inverse} - \text{Gamma}(1, 1)$$

For included coefficients,

$$\beta_z \mid \sigma^2 \sim \mathcal{N}(\mathbf{0}, g\sigma^2(X_z^T X_z)^{-1})$$

For excluded coefficients, $\beta_j = 0$.

We assume the uniform prior probability for all models, $p(\mathbf{z})$ is constant.

I will now use the Gibbs sampler to compute the posterior probability $p(\mathbf{z} \mid \mathbf{y}, \mathbf{X})$ via MCMC approximation.

$$\mathbf{z}^{(i)} \to (\sigma^2)^{(i)} \to \boldsymbol{\beta}^{(i)} \to \mathbf{z}^{(i+1)} \to \dots$$

1. Sample $\mathbf{z}^{(i+1)}$ (Model Selection): We update each z_j sequentially from its full conditional distribution, $P(z_j = 1 \mid \mathbf{y}, \mathbf{X}, \mathbf{z}_{-j}) = \frac{O_j}{1+O_j}$. The log odds ratio $r = \log(O_j)$ is computed using the log marginal likelihood function `lpy.X`.

2. Sample $(\sigma^2)^{(i+1)}$ and $\boldsymbol{\beta}^{(i+1)}$ (Model Averaging): We sample the parameters conditional on the newly selected model $\mathbf{z}^{(i+1)}$ using the `lm.gprior function`. These samples converge to the model-averaged posterior distribution $p(\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X})$.

The final model-averaged posterior mean $\hat{\boldsymbol{\beta}}_{MA}$ is approximated by the mean of the collected $\boldsymbol{\beta}$ samples: $\hat{\boldsymbol{\beta}}_{MA} = \frac{1}{S}\sum_{i=1}^{S} \boldsymbol{\beta}^{(i)}$.

I will use prior correspondence as follows:

$$a_z = \frac{1}{2}\nu_0 \implies \nu_0 = 2a_z; \quad b_z = \frac{1}{2}\nu_0 s_{20} \implies s_{20} = \frac{b_z}{a_z}$$

I will now proceed with this simulation, after importing functions from the given scripts.

```
set.seed(465)
source("scripts/regression_gprior.R")
az<-1->bz
BETA <- Z <- matrix(NA, S, p)
z <- rep(1, p)
lpy.c <- lpy.X(y, X[, z == 1, drop = FALSE],
               g = g, nu0 = 2*az, s20 = bz/az)

for(s in 1:S) {
    for(j in sample(1:p)) {
    zp <- z; zp[j] <- 1 - zp[j]
    lpy.p <- lpy.X(y, X[, zp == 1, drop = FALSE],
                   g = g, nu0 = 2*az, s20 = bz/az)
    r <- (lpy.p - lpy.c) * (-1) ^ (zp[j]==0)
    z[j]<-rbinom(1,1,1/(1+exp(-r)))
    if(z[j]==zp[j]){lpy.c<-lpy.p}
    }
  beta<-z
  if(sum(z)>0){beta[z==1]<-lm.gprior(y,X[,z==1,drop=FALSE],S=1,
                                     g = g, nu0 = 2*az, s20 = bz/az)$beta}
  Z[s,]<-z
  BETA[s,]<-beta
}
```

I will now compile the results into a neat tabular form.

```
credible_intervals <- t(apply(BETA, 2, quantile, probs = c(0.025, 0.975)))
results <- data.frame(
  Variable = var_names,
  Posterior_Prob = colMeans(Z),
  MA_Posterior_Mean = colMeans(BETA),
  MA_CI_Lower = credible_intervals[, 1],
  MA_CI_Upper = credible_intervals[, 2],
  CI_Excludes_Zero = (credible_intervals[, 1] > 0)|(credible_intervals[, 2] < 0)
)
results
```

```
##   Variable Posterior_Prob MA_Posterior_Mean MA_CI_Lower MA_CI_Upper
## 1                   1.0000       60.79760568  43.6821936  77.1028156
## 2    npreg          0.0932       -0.06067667  -0.9479639   0.0000000
## 3       bp          0.1574        0.03257227   0.0000000   0.3167871
## 4     skin          0.0986        0.02341563   0.0000000   0.3710826
## 5      bmi          0.9830        0.92253244   0.4223797   1.3355794
## 6      ped          0.6864        7.19306857   0.0000000  17.2722003
## 7      age          1.0000        0.73494569   0.4782514   1.0123738
##   CI_Excludes_Zero
## 1             TRUE
## 2            FALSE
## 3            FALSE
## 4            FALSE
## 5             TRUE
## 6            FALSE
## 7             TRUE
```

The model selection and averaging approach, which incorporates model uncertainty, yields a different and often more nuanced interpretation of which variables are truly predictive of glucose levels compared to the full Bayesian model from part (a).

The model averaging results provide two main pieces of information: the Posterior Inclusion Probability ($P(z_j = 1 \mid \mathbf{y}, \mathbf{X})$) and the model-averaged coefficients.

I can do model selection using the posterior probability. This column indicates the probability that a variable belongs in the model, after observing the data. Variables with a probability greater than 0.5 are strongly supported. The intercept, `age`, and `bmi` all have $P > 0.95$, indicating strong evidence exists that these variables are necessary predictors. `ped` has $P > 0.5$, i.e., moderate evidence exists that this predictor is necessary, since it is in 68% of all sampled models. The other predictors (`bp`, `npreg`, `skin`) do not show strong presence in most models. These variables are rarely included in the highly probable posterior models, suggesting they are likely irrelevant predictors.

The model-averaged posterior mean ($\hat{\boldsymbol{\beta}}_{MA}$) is the average coefficient across across all sampled models, weighted by their posterior probabilities. Based on the model-averaged credible intervals, only the intercept, `bmi`, and `age` are strongly predictive, as their 95% credible intervals do not contain zero. This confirms that the body mass index and age have a reliably positive relationship with glucose levels. The model-averaged posterior means for the weakly supported variables like `npreg`, `bp`, and `skin` are effectively shrink towards zero because they are set to zero in the majority of the sampled models (where $z_j = 0$). For instance, `npreg`'s model-averaged mean is -0.0607 compared to -0.6576 in the full model, showing that the model-averaged mean is closer to zero.

The key differences between the full model, which forces all seven variables to be in the model, and the model-averaged approach, which accounts for model uncertainty, are:

- The highly predictive variables where the 95% credible intervals does not cover 0, for the full model are `bmi`, `ped`, `age`, and the intercept. For the model-averaged version, these are `bmi`, `age`, and the intercept.
- The full model falsely identified diabetes pedigree as a strong predictor because it ignored model uncertainty.
- While the model-averaged version identified that `ped` has a moderate posterior probability, its model-averaged CI includes zero, indicating the evidence for a non-zero effect is not 95% certain once model uncertainty is accounted for.
- The Bayesian approaches correctly shrinks the coefficients for the variables close to zero, effectively eliminating them from the model, even if they were included.