STAT 578 – Advanced Bayesian Modeling – Fall 2019

Assignment 5

File `ozoneAQIaug.txt` contains daily ozone air quality index (AQI) values for fifteen different cities for the month of August 2018.[1] You will build and compare two different varying-coefficient hierarchical normal regression models for the *log*-values, using JAGS and `rjags`.

Let y_{ij} be the *natural logarithm* of the ozone AQI value for city j on day i ($i = 1, \ldots, 31$, $j = 1, \ldots, 15$), where the days are numbered as usual. For each city, model the log-value as a simple linear regression on the *centered* day number:

$$y_{ij} \mid \beta^{(j)}, \sigma_y^2, X \quad \sim \quad \text{indep. N}\big(\beta_1^{(j)} + \beta_2^{(j)}(x_i - \bar{x}),\, \sigma_y^2\big)$$

where

$$\beta^{(j)} \quad = \quad \begin{pmatrix} \beta_1^{(j)} \\ \beta_2^{(j)} \end{pmatrix} \qquad j = 1, \ldots, 15 \qquad\qquad x_i \;=\; i \qquad i = 1, \ldots, 31$$

Note that the coefficients are allowed to depend on the city, but the variance is not.

(a) Let $\hat{\beta}_1^{(j)}$ and $\hat{\beta}_2^{(j)}$ be the *ordinary least squares* estimates of $\beta_1^{(j)}$ and $\beta_2^{(j)}$, estimated for city j. (The coefficients are estimated completely separately for each city.)

  (i) [1 pt] Produce a scatterplot of the pairs $\big(\hat{\beta}_1^{(j)}, \hat{\beta}_2^{(j)}\big)$, $j = 1, \ldots, 15$.

  (ii) [1 pt] Compute the average (sample mean) of $\hat{\beta}_1^{(j)}$ and also of $\hat{\beta}_2^{(j)}$.

  (iii) [1 pt] Compute the sample variance of $\hat{\beta}_1^{(j)}$ and also of $\hat{\beta}_2^{(j)}$.

  (iv) [1 pt] Compute the sample correlation between $\hat{\beta}_1^{(j)}$ and $\hat{\beta}_2^{(j)}$.

(b) Consider the bivariate prior

$$\beta^{(j)} \mid \mu_\beta, \Sigma_\beta \quad \sim \quad \text{iid N}(\mu_\beta,\, \Sigma_\beta)$$

$$\mu_\beta \quad = \quad \begin{pmatrix} \mu_{\beta_1} \\ \mu_{\beta_2} \end{pmatrix} \qquad\qquad \Sigma_\beta \quad = \quad \begin{pmatrix} \sigma_{\beta_1}^2 & \rho\,\sigma_{\beta_1}\sigma_{\beta_2} \\ \rho\,\sigma_{\beta_1}\sigma_{\beta_2} & \sigma_{\beta_2}^2 \end{pmatrix}$$

with hyperpriors

$$\mu_\beta \quad \sim \quad \text{N}\big(0,\, 1000^2 I\big)$$

$$\Sigma_\beta^{-1} \quad \sim \quad \text{Wishart}_2\big(\Sigma_0^{-1}/2\big)$$

in the notation used in the lecture videos. For your analysis, use

$$\Sigma_0 \quad = \quad \begin{pmatrix} 0.1 & 0 \\ 0 & 0.001 \end{pmatrix}$$

based on preliminary analyses. Let the prior on σ_y^2 be

$$\sigma_y^2 \;\sim\; \text{Inv-gamma}(0.0001, 0.0001)$$

[1] Data from `https://www.epa.gov/outdoor-air-quality-data/air-quality-index-daily-values-report`

(i) [2 pts] List an appropriate JAGS model. Make sure to create nodes for Σ_β, ρ, and σ_y^2.

Remember that the ozone AQI values are to be analyzed on the *log* scale.

Now run your model using `rjags`. Make sure to use multiple chains with overdispersed starting points, check convergence, and monitor μ_β, Σ_β, σ_y^2, and ρ (after convergence) long enough to obtain effective sample sizes of at least 4000 for each parameter.

(ii) [2 pts] Display the `coda` summary of the results for the monitored parameters.

(iii) [2 pts] Give an approximate 95% central posterior credible interval for the correlation parameter ρ, and also produce a graph of its (estimated) posterior density.

(iv) [2 pts] Approximate the posterior probability that $\rho > 0$. Also, compute the Bayes factor favoring $\rho > 0$ versus $\rho < 0$. (You may use the fact that $\rho > 0$ and $\rho < 0$ have equal prior probability.) Describe the level of data evidence that $\rho > 0$.

(v) [1 pt] Your model implies that, over the 30-day period from the first day of August to the last, the (population) median ozone AQI value should have changed by a factor of

$$e^{30\mu_{\beta_2}}$$

Form an approximate 95% central posterior credible interval for this quantity.

(vi) [2 pts] Use the `rjags` function `dic.samples` to compute the effective number of parameters ("penalty") and Plummer's DIC ("Penalized deviance"). Use at least 100,000 iterations.

(c) Now consider a different model with "univariate" hyperpriors for the model coefficients, which do not allow for a coefficient correlation parameter:

$$\begin{aligned} \beta_1^{(j)} \mid \mu_{\beta_1}, \sigma_{\beta_1} &\sim \text{iid N}\big(\mu_{\beta_1},\, \sigma_{\beta_1}^2\big) \\ \beta_2^{(j)} \mid \mu_{\beta_2}, \sigma_{\beta_2} &\sim \text{iid N}\big(\mu_{\beta_2},\, \sigma_{\beta_2}^2\big) \end{aligned}$$

with hyperpriors

$$\begin{aligned} \mu_{\beta_1}, \mu_{\beta_2} &\sim \text{iid N}\big(0,\, 1000^2\big) \\ \sigma_{\beta_1}, \sigma_{\beta_2} &\sim \text{iid U}(0, 1000) \end{aligned}$$

Let the prior on σ_y^2 be the same as in the previous model.

(i) [4 pts] Draw a complete DAG for this new model.

(ii) [2 pts] List an appropriate JAGS model. Make sure that there are nodes for $\sigma_{\beta_1}^2$, $\sigma_{\beta_2}^2$, and σ_y^2.

Remember that the ozone AQI values are to be analyzed on the *log* scale.

Now run your model using `rjags`. Make sure to use multiple chains with overdispersed starting points, check convergence, and monitor μ_{β_1}, μ_{β_2}, $\sigma_{\beta_1}^2$, $\sigma_{\beta_2}^2$, σ_y^2 (after convergence) long enough to obtain effective sample sizes of at least 4000 for each parameter.

(iii) [2 pts] Display the `coda` summary of the results for the monitored parameters.

(iv) [2 pts] Recall the (population) median ozone AQI change factor

$$e^{30\mu_{\beta_2}}$$

considered in the previous analysis. Form an approximate 95% central posterior credible interval for this quantity, and compare it with the previous results.

(v) [2 pts] Use the `rjags` function `dic.samples` to compute the effective number of parameters ("penalty") and Plummer's DIC ("Penalized deviance"). Use at least 100,000 iterations.

(vi) [1 pt] Compare the (Plummer's) DIC values for this model and the previous one. Which is preferred?

(d) (i) [2 pts] It is possible that the variability in log-value depends on the city. How might you modify your model to account for this? Would your solution need more hyperparameters?

(ii) [1 pt] It is possible that there are time-series correlations in the successive log-values of each city that are not captured by the simple linear regression model. What specific model assumption would this violate? (Is i or j involved?)

(iii) [1 pt] It is possible that there are spatial correlations among the log-values on a given day that are not captured by the simple linear regression model (because some cities are closer to each other than others). What specific model assumption would this violate? (Is i or j involved?)

Total: 32 pts