Intermediate Bayesian Modeling

Monte Carlo Project

Due 13 December 2018 by 12pm (noon).

For this project, you will work alone or in groups of two, using the tools you've learned for generating random samples to understand a (made up) dataset. Details about the data set, as well as instructions for your work, will be provided below. You will be able to find the data for this project in the file below, which will be available on our class website:

`citations.csv`

This file contains the data for you to use in this project. It is structured as a comma-delimited data file, with each observation having its own line and values for variables separated by commas.

Background

The data set for this project is made up—and is not intended to suggest real-world phenomena, except in the most general sense. These made-up data are created to imitate citation count data for academic papers.

It is well known that citation counts for academic papers vary greatly from paper to paper, from journal to journal, and from academic discipline to discipline. It is often possible to model citation counts using a set of covariates—but are the important covariates the same for all papers? In this data set, I posit that the most model for citation counts may differ from paper to paper. This can be conceptualized in two different ways: as models with many interaction terms, and as a small number of models with a latent variable for each paper indicating which model it should use. For this project, we will consider the latter approach.

You will be given a set of three models for citation counts, along with a dataset generated according to these models and information for forming priors. Your job in this project will be to write code to generate a number of inferences based on these models, priors, and data; and to write a short report summarizing your findings.

Models and Variables

The following variables are provided for each paper:

`CITES`

Number of citations a paper has.

`AGE`

How long the paper has been in print (measured in years).

`JOURNAL`

The impact factor (IF) of the journal in which the paper was published.

`PAGES`

Number of pages in the paper.

`REFERENCES`

Number of papers referenced by this paper.

`AUTHORS`

Number of authors on the paper.

`FUNDING`

Indicator variable for whether the authors receive funding in support of their research, from either industry or government (0 if research was unfunded, 1 if research was funded).

As count data, citation counts can be expected to follow poisson distributions. Each of the three models gives a different poisson parameter for these counts: λ_i, $i \in \{1, 2, 3\}$.

$$\begin{aligned}
\lambda_1 &= \exp\left(\beta_0 + \beta_1 X_{age} + \beta_2 \log X_{journal}\right) \\
\lambda_2 &= \exp\left(\gamma_0 + \gamma_1 X_{age} + \gamma_2 \log X_{journal} + \gamma_3 X_{pages} + \gamma_4 X_{references}\right) \\
\lambda_3 &= \exp\left(\eta_0 + \eta_1 X_{age} + \eta_2 \log X_{journal} + \eta_3 X_{authors} + \eta_4 X_{funding}\right)
\end{aligned}$$

As a personal note, I'm generating these models based on my own experience working in statistics, psychology, and biostatistics. The first model is most typical of discussions I've had with mathematical sciences faculty, where papers are largely cited based on their own contributions. The second model is more typical of discussions I've had about meta-analyses—which are common in psychology and medicine—where a great number of papers are cited and information from those papers is combined to provide an overall picture of the state of knowledge in a particular area. The third model is more typical of discussions I've had with collaborators in the medical sciences, where large projects with funding tend to be among the most important publications.

Instructions

The report for this project will be formatted as answers to a number of questions provided here to guide you through the analysis and simulation. If you are working in a group of two, one student should answer questions (1) and (2) and the other should answer questions (3) and (4). Work together to accomplish question (5). Grading will be more difficult for groups of two.

(1) To allow each paper to associate with the most appropriate model, you will need to introduce a latent variable z_i indicating which model is appropriate for each paper. For each paper i, this latent variable should follow a $z_i \sim \text{Mult}(1, p)$ distribution where p is a vector of probabilities for each model.

   a. What type of prior should you place on p? Use a reference prior for this parameter—assume you don't have useful prior information on the proportion of papers that are likely to follow the various models.

   b. Explain how to generate a random sample from this prior using only independent Uniform $(0, 1)$ random variables (or if you can't, explain why you can't).

c. Try it—use independent uniforms to generate a random sample ($n = 10,000$) from your prior or a similar prior you can simulate from in this way. Then generate a random sample ($n = 10,000$) using a standard random number generation function for the distribution. How do the means of these two samples compare? How do their covariance matrices compare?

(2) Write out the full likelihood equation for these data. You would like to find an expression for $L(\theta \mid Y)$ where $Y = \{y_1, ..., y_n\}$ are the observed citation counts. In this case $\theta = (p, \beta, \gamma, \eta)$ where we assume β, γ, and η refer to the coefficient vectors for models 1, 2, and 3 respectively. As we've discussed in class, a marginalized form with just $L(\theta \mid Y)$ is ideal—but it will probably be impossible for you to find a convenient one. It will be much easier to write the likelihood taking into account the additional latent variable $Z = \{z_1, ..., z_n\}$, call it $L(\theta \mid Y, Z)$. This is fine, and should be your goal.

You should expect finding the likelihood to be fairly difficult, and it deserves a considerable amount of your time. The likelihood does not need to be written on one line—for example, it will be much easier if you use λ_1, λ_2, and λ_3 in the equation and then explain once again what these λ's are. Remember also that the best way to combine multiple models when only one can be active at a time is to use the latent z_i's as an exponential "switch": turning on the parts of the equation that are active and turning off the parts that are not, much like the Bernoulli distribution.

(3) Again, since this is a difficult problem, prior information on the model parameters is hard to come by. Using standard reference priors for the model parameters in generalized linear models can lead to poor convergence, however—so you'll want to be careful in trying to decide on priors for these parameters. Think about what citation counts would be reasonable (based on what you know about citations or can learn from sources other than these data), and use this information to form an heuristic about what values of λ are reasonable. Then, log-transform your reasonable values for λ and use them as a guide for figuring out what range of support makes sense for your model parameter priors. Make sure to account for the variability in the covariates as well: you want to keep the range of possible $X\beta$ values under control, which means if X has large values, β will need to have correspondingly small values. Present the priors you choose and discuss how you arrived at them. Use independence priors for this problem—it's hard enough already without worrying about dependency between priors.

(4) Write out the joint posterior distribution of all the parameters, up to a proportionality constant. This is just the standard *[likelihood]* $\times$ *[prior]* formulation we're used to, but it will be unusually complicated because of all the parameters in this model. Be careful to include the probability model for the latent variables as well. Use the joint posterior to find the full conditional distributions for each parameter (again up to a proportionality constant). Again, be careful to correctly account for the latent data and the probability model associated with it.

(5) Construct a posterior sample for these data using the techniques you've learned in lectures and from Chapter 6 of our textbook. Report standard posterior inferences for the parameters (e.g. means, variances, 95% probability intervals).