A step by step guide to performing a Bayesian analysis on a data set

Andrew Parnell

1. **Exploratory data analysis (EDA)**. Plot your data to explore relationships, look for odd/missing values, and make sure you understand what you are analysing. Write down the key questions that you want to answer for this data set.
2. **Define variables**. Choose which variables are your response(s), which are continuous covariates, which are categorical covariates. Think about mean correcting or standardising continuous covariates.
3. **Initial model building**. Pick a very simple model (maybe a regression model/GLM, maybe something else, depending on your data) and draw a DAG. From the DAG identify which variables will appear in your likelihood and which will require priors.
4. **Find prior distributions**. Either from previously published work, or expertise of the data collectors/scientists behind the data set, try and create informative prior distributions for the relevant parameters. Use the prior predictive distribution to help guide you in this choice. If you have no information, try and put a prior distribution that at least matches the likely range of the parameter. If this is impossible put a vague prior on the parameter (e.g. $N(0, 1000)$).
5. **Run a first model**. Fit your first model in JAGS or Stan. Check for convergence using Rhat. These should all be <1.1. Plot and print the output and try to interpret the parameter values. Record the DIC/WAIC for this model, and the effective number of parameters, and check it roughly matches (or is less than) the number of parameters you are putting priors on in the model.
6. **Model expansion**. Try to make your model richer by including all the covariates you think are important, and putting prior distributions on them that borrow strength between groups. Draw a DAG for each model you create. It's best if you can build a single model that contains everything, and has priors such that your simple model is a special case of the new model. However, sometimes you will need to fit separate models and compare them using e.g. DIC. If possible, make sure there are parameters in the model which directly answer the questions of interest.
7. **Model checking**. With your final fitted model, make really sure it has converged (perhaps do a longer run over night) and produce a posterior predictive check. Also check other assumptions in the model, such as residual or outlier analysis. If precise prediction is a big concern, consider running a 5 or 10-fold cross validation.
8. **Model summaries**. Summarise the output of your model through plots and tables. Try to use all of the posterior samples, for example by plotting fitted lines with confidence intervals, or by creating density plots or histograms of the posterior distribution of the parameters of interest. If you want to try and create a Bayesian p-value (do you really?) then, for example, find the proportion of posterior samples that are bigger (or less than) 0.