6.882 Presentation: Visualization in Bayesian Workflow

Tarek Mansour, Edgar Minasyan, Chandler Squires

Part 1 (Chandler): Overview and Model Choice

Throughout this course, we've developed a powerful toolbox. We've looked at modeling techniques like topic models and nonparametrics, and inference algorithms like ADVI and Hamiltonian Monte Carlo. While these are great in theory, as we've probably seen throughout the course of working on our projects, the hard part is putting these techniques into practice. To help us out with that, we've looked at a set of tools for evaluation.

Last week, we talked about Simulation-Based Calibration, which provides a fundamental snanity check of whether or not our model is implemented correctly. Then on Thursday, we talked about some problems particular to MCMC samplers, like getting decent mixing, and we talked about some diagnostics we can do to fix any problems with mixing. Today, we come to the last topic in this section, and this class, which is a super useful one now that we're going out into the world to actually use these techniques, and that's actually coming up with a model and improving it until we think it explains our data. This paper outlines a workflow for this and gives us some visual tools we can use to look for places to improve our model.

The approach the authors advocate is an iterative one of model extension, where we start with a simple model and make it more and more sophisticated until we're happy with how it explains the data. The

- Choosing a model (data to use, dependencies between variables, distributions)
- Choosing prior (vague vs. weakly informative)
- Checking algorithm (even if implemented correctly, could have some trouble, e.g. the funnel of death)
- Checking posterior (make sure it kinda looks like the data)

The example that they use throughout the paper is this problem of predicting pollution level in areas without the sensors that can measure it, using this technique called regression calibration, which in general takes some variable that's related to the variable of interest but more widely available and finds the relationship between them given the data available. In this case, the variable that they use for calibration is satellite measurements of aerosol optical depth, which measures how much aerosols in the atmosphere block light by absorption or scattering.

The first model they start out with is a simple linear regression between the log of the particulate matter concentration and the log of the aerosol optical depth. Even here a lot of choices are already being made: we're ignoring all of the information about location and picking linear regression on their log scales, rather than linear regression on the original scale or a more sophisticated regression like GPR.

Show plots using other regressions and final regression

When they think about ways to expand this model, they make an offhand comment about Simpson's paradox. This is actually a really important point because its connected to one of the main reasons that we would want to expand a model: there are some confounding variables we're not taking into account. The example of Simpson's paradox that I've always seen is in the context of medical treatments and I thought it would be nice to walk through that and give some perspective then link it back to the current problem.

* draw medical treatment example on board*

* draw regression example on board*

Here's where they start considering the location information from the data. I guess partially because it's easier to show, but also because they don't want to divide up the data too much, they use the super-region information rather than country or city. Now when they do the regression, they see some variation in the slopes around the slope they got when using all the data together. So the next thing they come up with is a model where there's a baseline slope and intercept and there's some variation around that.

* draw original model and hierarchical model on board*

And finally they give a third model that gives us an idea for an alternative when we don't have labels, and actually ends up performing better when we get to the section on posterior predictive checks. Instead of using labels, we can essentially assign them ourselves by running some sort of clustering like K-means, which would work really well for the Simpson's paradox example. What they actually do is group all the ground monitor measurements from a country together into a unit, then cluster the countries based on their average particulate matter concentration. They choose to break the data up into 6 clusters, I think mostly to compare with the model 2.