Stat238: Problem Set 2
Due Wednesday Oct. 5

September 19, 2016

Comments:

- Please note my comments in the syllabus about academic integrity. You can work together on the problems but your final writeup must be your own and you should wrestle with the problems on your own first. A couple of the problems are from lab; overlap with your groupmates on those problems is fine.
- It's due at the start of class on paper. The syllabus discusses the penalty for turning it in late.
- Please give the names of anyone you worked with on the problem set on what you hand in.

Problems

1. Lab 4 problem. Consider the following data on the number of fatal accidents on scheduled airline flights per year. Assume that the number of accidents in year t follows a Poisson distribution with mean $\alpha + \beta t$ and that the number of accidents is conditionally independent between years, given α and β.

Year	'76	'77	'78	'79	'80	'81	'82	'83	'84	'85
Accidents	24	25	31	31	22	21	26	20	16	22

   (a) Choose a non-informative prior distribution for (α, β).

   (b) Write the joint posterior density for (α, β). Do you recognize any known distributional form for $\alpha|\beta, y$ or $\beta|\alpha, y$?

   (c) Calculate crude estimates and uncertainties for (α, β) using a non-Bayesian analysis.

   (d) Plot the contours of the joint posterior density. If you're doing this in R, the *contour()* function should work - see *help(contour)*, which will indicate that you need the posterior calculated on a grid (see *expand.grid()*), but then need the values in the form of a matrix.

   (e) Draw 1000 random samples from the joint posterior density of (α, β). Overlay them on the contour plot. Do things seem consistent?

   (f) Create simulation draws and plot a histogram of the posterior predictive distribution for the *number* of fatal accidents in 1986. Compute a 95% posterior predictive credible interval.

   (g) Find and plot the 95% highest posterior density (HPD) region for (α, β) using the discrete approximation.

(h) Consider your discrete approximation. How many calculations of the posterior density did you do? How many would you have had to do if the parameter vector had 10 parameters instead of two? How long do you project it would take (roughly) to do the computation in 10 dimensions and how big would a file containing the results be (each number requires 8 bytes). This is the curse of dimensionality in the context of Bayesian computation.

(i) Use the Bayesian central limit theorem result to determine a normal approximation to the posterior and compare to the true posterior based on comparing contour plots (ideally overlaid on top of each other). How well does the approximation work in this case?

2. Write out the likelihood for the problem posed in BDA 3.5. You don't need to do any of the rest of the problem. Side note: this is a real problem – I'm currently working on a more complicated version of this problem with some global health researchers who have blood pressure data where the measurements are either rounded to the nearest 10, 5, 2, or 1, but for any given measurement, we don't know what kind of rounding has been used!

3. Practice with the multivariate normal: Consider a basic nonparametric regression using splines:

$$Y \sim N(Bu, \sigma^2 I)$$

where Y is an n-vector of observations, B is a fixed spline basis matrix and the k-length vector u is the basis coefficients. Penalized splines involve a penalty on u, and in a Bayesian formulation, the prior plays the role of the penalty. The simplest approach is to take $u \sim N_k(0, \tau^2 I)$, discussed in Ruppert et al.'s book *Semiparametric Regression*, but others argue for other types of penalties. Let's retain generality and assume $u \sim N_k(\mu_0, P)$ where P is an arbitrary covariance matrix. Derive the conditional posterior, $p(u|Y, \sigma^2, P, \mu_0)$. Note that you can solve this problem as we did in class without fully completing the square. (If you're not familiar with splines, you don't need to know anything about them for this problem, but feel free to ask me as they're a useful way of estimating non-linear regression relationships.)

4. Suppose we use a uniform prior for a standard deviation parameter, $\tau \sim U(0, \infty)$. What is the induced prior on τ^2? Express this as a degenerate form (i.e., it's not integrable) of a distribution whose functional form we are familiar with.

5. BDA 5.10a and 5.10b. Note that this involves some real analysis. If you're coming from a more applied statistics background, do your best, but one problem won't make or break your grade.
Hints:

(a) for 5.10a, consider the behavior of $p(\tau|y)$ for values of τ near 0 and think about putting a bound on all the terms in $p(\tau|y)$ other than $p(\tau)$.

(b) for 5.10b, try to put an upper bound on $p(\tau|y)$ for τ near zero and for τ away from 0 and show that both parts are integrable.

6. BDA 5.15. We'll likely have this as part of a lab, and I may extend the problem a bit so consider this question to still be in draft form.