CSC 2541, Assignment #2, due in class on November 24. Worth 16% of the mark.

For this assignment, you will fit Bayesian mixture models with five components and with an infinite number of components to some four-dimensional artificial data, using the mixture modeling facilities of my software for flexible Bayesian modeling.

The data is available from the web page, and also in /u/radford/course/csc2541/ass2-train and /u/radford/course/csc2541/ass2-test on the CSLAB and UTSTAT systems. There are 1000 training cases and 500 test cases. They are stored one per line, with each line containing four numbers.

You should model this data using a Bayesian mixture model, with Gaussian components, using priors that treat each variable the same (ie, assume that you have no prior knowledge that distinguishes one variable from another). You should allow the four variables to have different hyperparameter values, however — ie, the variables don't have to look the same in the posterior distribution.

You should try models with five components and with a countably infinite number of components, and compare how well they perform in terms of predicting the test cases (ie, in terms of average log predictive probability density over test cases, found using a sample from the posterior distribution). You should also see how sensitive the performance is to the setting of the concentration parameter for the Dirichlet process (α), trying a value of one and some higher values.

You should also use your best model to find the conditional distibution for the first variable given the other three for each of the test cases. To do this for a test case, you will need to evaluate the predictive density for a number of cases that have the same values as the test case for all but the first variable, but with the first variable set to a grid of possible values. After normalizing these probabilities to sum to one, they will form an approximation to the conditional distribution. Comment on how the forms of these distributions vary from test case to test case.

Hand in any command scripts and programs you used, along with some informative output and plots, and your discussion.