Gibbs Sampling with the MDP Model

In the lectures, we’ve discussed a basic Gibbs sampler for the MDP model and how to modify it to improve convergence and mixing. We’ll focus on a simple model and a stylized estimation problem. The Gibbs sampler we’ll use is that of Bush and MacEachern (1996).

The web site contains an R workspace and a script with R code to fit the model. The small R workspace contains only data—I recommend downloading this one, and only moving to the large R workspace if you have trouble with the code. The two data files are on weights of men. The first, bodyfat, is from the StatLib archive housed at Carnegie Mellon University. The R script contains a pointer to StatLib. The second is from a roster of the 2008-2009 Buckeye football team.

The model is the basic MDP-DPM mixture of normals, where, for simplicity, all components of variation are assumed known. We will use it for the compound decision problem, although in this setting, the goal of estimating $\theta_i$ is artificial.

Functions

The requisite generations for the Gibbs sampler consist of

1. Generate $\theta_i$
   
   (a) Remove $\theta_i$ from the cluster structure
   
   (b) Generate a new value for $\theta_i$

2. Generate the $\theta^*$

3. Generate $\mu$

In addition, there are two functions to control the Gibbs sampler.

1. Run one iterate of the Gibbs sampler
2. Run a collection of iterates (and collect output)

Output

Once you have the Gibbs sampler up and running, there is extraordinary variety to the output that you may wish to collect and store. The functions above preserve a tiny bit of Markov chain as output. With thought and planning (i.e., what will you look at for diagnostic purposes? for the MCMC? for the model? What are the primary features of inference?), more output should be stored. For this model, one might choose to store the number of clusters, $k$, the clustering vector, $s$, and even perhaps the mean of the distribution from which each $\theta^*$ is generated, immediately preceding its generation.
Figure 1: Time series plots of $\mu$ (upper left panel), and the first fifteen thetas for the body fat data set. The thetas proceed across rows, then down columns.
Convergence diagnostics

Convergence diagnostics are essential to soundly fitting a model via MCMC. The basic set of diagnostics track individual parameters across iterates of the Markov chain. When making these plots, plot the individually generated values as points, not as lines connecting successive points. The first two figures provide time series plots of several parameters that should be checked when assessing the performance of the Markov chain.

Like most of Statistics, diagnostics are an art form. The goal of making plots is to have a clear focus for the plot (recall the use of different residuals for the various residual plots in a regression analysis), and to make them as informative as possible.

To many, the main message of ANOVA is the decomposition of total variation into a set of component parts. How does this connect to diagnostic plots? When a value of $\theta_i^*$ is generated, it becomes the new value for each of the $\theta_j$ attached to it. One component of variation accounts for differences in the mean of the full conditional posterior for $\theta_i^*$. A second component of variation handles the generated value of $\theta_i^*$. Removing this second component of variation will often produce more informative plots. Unfortunately, a more informative plot will not suggest better convergence/mixing, but it may suggest poorer convergence/mixing. Drawing an analogy with hypothesis testing, use of the standard plot is much like choosing to use a hypothesis test with low power–because life will be much more pleasant if the null hypothesis test is not rejected.
Figure 3: Plots of estimate (posterior mean of $\theta_i$) against $X_i$ for the body fat data.

Estimation

The purpose of forming the Markov chain is merely to fit the model. It gives insight into the posterior distribution in two main ways: It allows us to explore many features of the posterior distribution—what does the posterior density of $\theta_1$ look like? of $\mu$? How many clusters do there appear to be? etc. It also allows us to compute estimates of a variety of quantities. Most often, we use the posterior mean for our estimates. This seems to be as much a matter of ease as of principle. The last figure provides plots of these estimates against the data values. The top row is for the hierarchical MDP model; the bottom row for the non-hierarchical version of the model. The left column is for the body fat data, the right column for the Buckeye data.