Summary of “The Infinite Gaussian Mixture Model” (Rasmussen, 2000)

Matt Dickenson `mcd31`

STA571/CS590.01 Due: 13 January, 2014

The use of (Gaussian) mixture models is valuable for the analysis of data which is thought to have originated from one of several latent components. However, because the indicators identifying the component from which each observation originated are “missing data,” it can be difficult to choose the number of components k. Because k is unknown, it is useful to allow it to vary infinitely. This paper introduces a practical method for estimating infinite hierarchical Bayesian mixture models. To do so, it combines two threads of previous research: the extension of finite Gaussians to the limiting case, and Dirichlet process mixture models. The result is that infinite Gaussian mixture models can be estimated by MCMC in finite time.

One key “trick” in the paper is to substitute k_{rep}, the number of classes that (currently) have data associated with them, for the parameter k when computing conditional posteriors for all model variables except the indicators (p. 4). That is, within the infinite possible values of k, there are some represented and some unrepresented, and the “true” value may be in either category. The parameter distributions for every k in the unrepresented class are all the same (p. 4). At each iteration of the MCMC, we can sample from the priors to get a Monte Carlo estimate of generating a new class. The paper demonstrates that infinite Gaussian mixture models can obtain good performance on multidimensional data without overfitting.