AMS 241: Bayesian Nonparametric Methods (Spring 2018)

Semiparametric and nonparametric regression with DP mixture models
(due Thursday June 7)

Note. The requirement for this homework assignment is to submit a solution for one of the two problems. Working on both problems is optional (it will give you extra credit).

1. Consider data on the incidence of faults in the manufacturing of rolls of fabric:

`http://www.stat.columbia.edu/~gelman/book/data/fabric.asc`

where the first column contains the length of each roll, which is the covariate with values x_i, and the second column contains the number of faults, which is the response with values y_i, for $i = 1, ..., n$, with $n = 32$.

A Poisson regression is a possible model for such data, where the y_i are assumed to arise independently, given parameters $\theta > 0$ and $\beta \in \mathbb{R}$, from Poisson distributions with means $\mathrm{E}(y_i \mid \beta, \theta) = \theta \exp(\beta x_i)$, such that $\log(\theta)$ is the intercept and β is the slope of a linear regression function under a logarithmic transformation of the Poisson means. The Bayesian model is completed with priors for θ and β.

The Poisson regression can be extended in a hierarchical fashion to allow for over-dispersion relative to the Poisson response distribution. In particular, the response distribution can be extended to a negative Binomial under the following hierarchical structure:

$$\begin{aligned} y_i \mid \theta_i, \beta &\overset{ind.}{\sim} \text{Poisson}(y_i \mid \theta_i \exp(\beta x_i)), \quad i = 1, ..., n \\ \theta_i \mid \mu, \zeta &\overset{i.i.d.}{\sim} \text{gamma}(\zeta, \zeta\mu^{-1}), \quad i = 1, ..., n \end{aligned}$$

such that the mean of the gamma distribution for the θ_i is μ and the variance is μ^2/ζ. Under this hierarchical model, $\mathrm{E}(y_i \mid \beta, \mu, \zeta) = \mu \exp(\beta x_i)$ and $\mathrm{Var}(y_i \mid \beta, \mu, \zeta) > \mu \exp(\beta x_i)$, thus achieving over-dispersion relative to the Poisson regression model. In this case, the Bayesian model is completed with priors for β, μ and ζ.

Develop a semiparametric DP mixture regression model for the count responses y_i, which includes as limiting cases both of the parametric regression models discussed above. Discuss prior specification for your DP mixture model, and implement it for the specific data set (you can use any MCMC algorithm you wish, but you should write your own code). Compare the inference for the mean regression function arising from the two parametric models and from the semiparametric DP-based extension. Use a model comparison criterion for more formal comparison of the three models.

2. Consider the data set `ozone` from the "ElemStatLearn" R package. The data set includes measurements of ozone concentration in parts per billion, wind speed in miles per hour, daily maximum temperature in degrees Fahrenheit, and solar radiation in langleys, recorded over 111 days from May to September of 1973 in New York.

Develop a DP mixture of multivariate normals model for the joint distribution of the four variables included in this problem, mixing on both the kernel mean vector and covariance matrix. Discuss prior specification for your DP mixture model. Implement the model (again, you can use any MCMC algorithm you wish, but write your own code) and develop inference for bivariate densities for specific pairs of variables, as well as inference for conditional relationships between the variables. You can consider ozone concentration as the response variable, but also explore further conditional relationships of interest.