AMS 241 - Project: Density Estimation with DP Mixture

Tracy Holsclaw

March 2009

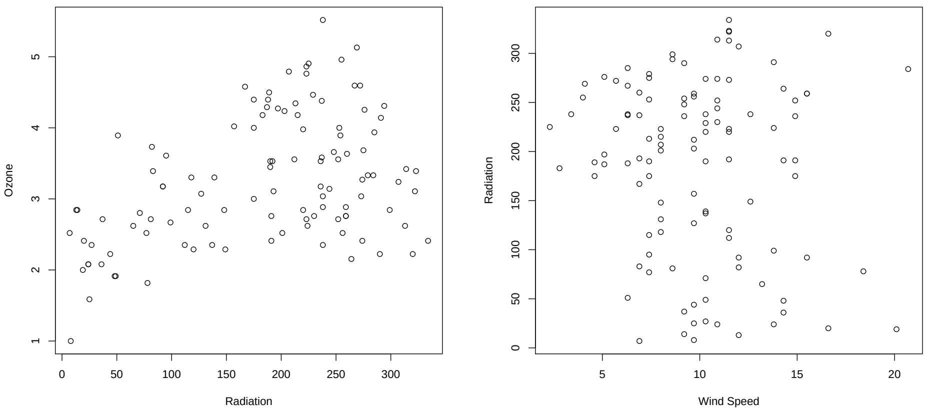
1 Introduction

The goal of this project will be to explore density estimation and regression fitting with a Bayesian non-parametric approach. This type of model will easily accommodate multimodal distributions and lead to posterior predictive density plots over a grid, conditional plots, and regression type fits and probability intervals. Two different covariate and response sets were chosen to display different attributes of these types of plots and analysis that can be done under this model. The model will be described and then an algorithm implemented using a Blocked Gibbs approach and advantages of this method are discussed. The data here is a simple example but this approach can be extended to many other more complex data structures that will not be discussed.

2 Density Model

We will use two data sets for our example both with n=111. This is a stock example from R: in this analysis, however, Ozone will be transformed to Ozone$^{\frac{1}{3}}$. We are looking at one covariate and one response in two separate set ups both of which are continuous. The clustering is different for these two set ups which will lead to interesting density, regression, and conditional density plots.

We would like to do density estimation over both data plots and also fit a regression type line of $E(y|x)$. We are assuming a non-parametric form is best for the density estimation because the data plots show signs of multimodality and clustering and we wish our model to reflect this. A non-parametric approach to regression will also be employed because we do not believe this to be a linear relationship and do not want to specify a parametric form for our curve.



(a) Radiation vs. Ozone

(b) Wind vs Radiation

Figure 1: Data

The model we will be:

$$
\begin{aligned}
(x_i, y_i)|\mu_i, \Sigma_i &\sim MVN((x_i, y_i)|\mu_i, \Sigma_i) && (1)\\
\mu_i, \Sigma_i|G &\sim G, \quad i = 1, ..., n && (2)\\
G|... &\sim DP(\alpha, G_0 = MVN(\mu_i|\mathbf{a}_\mu, B_\mu)IW(\Sigma_i|a_\Sigma = 4, B_\Sigma)) && (3)\\
\alpha, \mathbf{a}_\mu, B_\mu, B_\Sigma &\sim \pi(\alpha)\pi(\mathbf{a}_\mu)\pi(B_\mu)\pi(B_\Sigma) && (4)
\end{aligned}
$$

The priors are:

$$
\begin{aligned}
\pi(\alpha) &\sim \Gamma(shape = a_\alpha, rate = b_\alpha) && (5)\\
\pi(\mathbf{a}_\mu) &\sim MVN(\mathbf{a_1}, B_1) && (6)\\
\pi(B_\mu) &\sim IW(a_3 = 4, B_3) && (7)\\
\pi(B_\Sigma) &\sim W(a_4 = 4, B_4) && (8)
\end{aligned}
$$

3 Blocked Gibbs Sampler

There are several ways to proceed with this model. One could use a standard Gibbs sampler but there are issues with $q_0 = \alpha \int MVN(y_i|\theta)MVN(\theta|\psi)IW(\theta|\psi)d\theta$ not being analytically solvable in this case. There are two obvious choices on how to handle q_0 in this case: one could do numerical integration of q_0 or it could be approximated by $q_0 = \alpha \int MVN(y_i|\theta)MVN(\theta|\psi)d\theta$ as like in (Muller et al., 1996). Neither of these options is terribly satisfying as they are both approximate solutions and thankfully there are better options. One of these other options includes marginalizing over G and using an algorithm like the ones presented in Neal 2000. This is a good option but extra work is needed in the end to create regression curves. Instead, we will opt to use blocked Gibbs sampler. This is based on stick-breaking with truncation the process at N instead of letting it go to infinity. $\mathbf{p} = (p_1, ..., p_N)$ and $\theta_i = (\mu_i, \Sigma_i)$ and we will denote $\mathbf{y}_i = (x_i, y_i)$. We start with:

$$\mathbf{y}_i|\theta_i \sim MVN(\mathbf{y}_i|\theta_i) \tag{9}$$
$$\theta_i|G \sim G, i = 1, ..., n \tag{10}$$
$$G|\alpha, \psi \sim DP(\alpha, G_0(\psi)) \tag{11}$$

After doing the truncation at N; we use $G_N = (\mathbf{p}, \mathbf{Z})$ in place of G with $\mathbf{Z} = (Z_1, ..., Z_N)$.

$$\mathbf{y}_i|\theta_i \sim k(\mathbf{y}_i|\theta_i) \tag{12}$$
$$\theta_i|\mathbf{p}, \mathbf{Z} \sim G_N, i = 1, ..., n \tag{13}$$
$$\mathbf{p}, \mathbf{Z}|\alpha, \psi \sim f(\mathbf{p}|a) \prod_{l=1}^{N} g_0(Z_l|\psi) \tag{14}$$
$$\alpha, \psi \sim \pi(\psi)\pi(\alpha) \tag{15}$$

Other needed equations and facts for this stick-breaking construction:

$$V_l \sim Beta(1, \alpha), l = 1, ..., N-1 \tag{16}$$
$$p_1 = V_1, p_l = V_l \prod_{r=1}^{l-1}(1 - V_r) for l = 2, ..., N-2, p_N = \prod_{r=1}^{N-1}(1 - V_r) \tag{17}$$
$$f(\mathbf{p}|\alpha) = \alpha^{N-1} p_N^{\alpha-1}(1 - p_1)^{-1}(1 - (p_1 + p_2))^{-1} * ... * (1 - \sum_{l=1}^{N-2} p_l)^{-1} \tag{18}$$

Next we marginalize over θ_i to come to the blocked Gibbs sampler set up.

$$\mathbf{y}_i|\mathbf{z}_i, L_i \sim MVN(\mathbf{y}_i|Z_{L_i}) \tag{19}$$
$$L_i|\mathbf{p} \sim \sum_{l=1}^{N} p_l \delta_l(L_i), i = 1, ..., n \tag{20}$$
$$\mathbf{p}|\alpha \sim f(\mathbf{p}|a) \tag{21}$$
$$Z_l|\psi \sim G_0(.|\psi), l = 1, ..., N \tag{22}$$
$$\alpha, \psi \sim \pi(\psi)\pi(\alpha) \tag{23}$$

3.1 Notation

- n^* is the number of clusters $j = 1, ..., n^*$
- L_i (w_i) is a vector of indexes corresponding to which cluster $Z_{L_{j^*}}$ (θ_j^*) belongs
- There are n_j, $Z_{L_{j^*}}$ in each cluster
- There is one Z_l for $l = 1, ..., N$
- $Z_{l,(1,2,3)} = (\mu_1, \mu_2, \Sigma)_l$

4 Posteriors

4.1 Update z_l

Remembering that:
$g_0 = MVN(\mu_i|\mathbf{a}_\mu, B_\mu)IW(\Sigma_i|a_\Sigma = 4, B_\Sigma)$
$p(Z_l|...) \sim g_0(Z_l|\psi) \prod_{j=1}^{n^*} \prod_{i:L_i=L_j^*} MVN(\mathbf{y}_i|Z_{L_{j^*}})$

1. If l is not in $\{L_{j^*} : j = 1, ...n^*\}$ then we draw a new $Z_{l,(1,2,3)}$ from the prior (as a consolation this prior is being update by its hyperparameters). Basically, if the current cluster has no data currently assigned to it then we want redraw this cluster value from the prior.
$Z_{l,(1,2)}|Z_{l,(3)} \sim MVN(\mathbf{a}_\mu, B_\mu)$
$Z_{l,3}|Z_{l,(1,2)} \sim IW(a_\Sigma, B_\Sigma)$

2. If l is in $\{L_{j^*} : j = 1, ...n^*\}$ then this cluster is being favored by the data. We do not want to start over and draw the new parameter values from the prior but rather just want to update the values slightly.

$$Z_{l,(1,2,3)} \sim MVN(Z_{l,(1,2)}|\mathbf{a}_\mu, B_\mu)IW(Z_{l,3}|a_\Sigma = 4, B_\Sigma) \prod_{i:L_i=L_j^*} MVN(\mathbf{y}_i|Z_{L_{j^*}}) \tag{24}$$

$$Z_{l,(1,2)}|Z_{l,(3)} \sim MVN\left((B_\mu + n_j\Sigma^{-1})^{-1}(B_\mu^{-1}a_\mu + Z_{l,3}\sum_{i=1}^{L_i} y_i), (B_\mu + n_j\Sigma^{-1})^{-1}\right) \tag{25}$$

$$Z_{l,3}|Z_{l,(1,2)} \sim IW(a_\Sigma + n_j, B_\Sigma + (\mathbf{y}_i - Z_{l,(1,2)})'(\mathbf{y}_i - Z_{l,(1,2)})) \tag{26}$$

4.2 Update p_l and L_i

$$
\begin{aligned}
&M_l = \{i : L_i = l\} && (27)\\
&V_l^* \sim Beta(1 + M_l, \alpha + \sum_{r=l+1}^{N} M_r) && (28)\\
&p_1 = V_1^*, p_l = V_l^* \prod_{r=1}^{l-1}(1 - V_r^*) forl = 2, ..., N-1, p_N = 1 - \sum_{l=1}^{N-1} p_l && (29)\\
&L_i = Discrete(p_l MVN(\mathbf{y}_i|Z_l)) && (30)
\end{aligned}
$$

4.3 Update α

$$
\begin{aligned}
\alpha|... &\propto \Gamma(a_\alpha, b_\alpha)\alpha^{N-1}p_N^\alpha && (31)\\
&\sim \Gamma(N + a_\alpha - 1, b_\alpha - \ln p_N) && (32)\\
&\sim \Gamma(N + a_\alpha - 1, b_\alpha - \sum_{r=1}^{N-1} \ln(1 - V_r^*)) && (33)
\end{aligned}
$$

4.4 Update hyperparameters

I am tired of typing so here are the posteriors.

$$
\begin{aligned}
p(\mathbf{a}_\mu, B_\mu, B_\Sigma) \propto \pi(\mathbf{a}_\mu, B_\mu, B_\Sigma) \prod_{j=1}^{n^*} g_0(Z_{L_{j^*}}|\mathbf{a}_\mu, B_\mu, B_\Sigma) && (34)\\
p(\mathbf{a}_\mu|...) \sim MVN\left((B_1^{-1} + n^* B_\mu^{-1})^{-1}(B_1^{-1}\mathbf{a_1} + B_\mu^{-1}\sum Z_{l,(1,2)})^{-1}, (B_1^{-1} + n^* B_\mu^{-1})^{-1}\right) && (35)\\
p(B_\mu|...) \sim IW\left(a_3 + n^*, B_3 + \sum_{j=1}^{n^*}(Z_{l,(1,2)} - \mathbf{a}_\mu)'(Z_{l,(1,2)} - \mathbf{a}_\mu)\right) && (36)\\
p(B_\Sigma|...) \sim W\left(a_4 + n^* a_\Sigma, (B_4^{-1} + \sum_{j=1} n^* Z_l^{-1})^{-1}\right) && (37)
\end{aligned}
$$

5 Prior specification

$\alpha \sim \Gamma(5, 1)$; I tried a few things and this worked well. We are using a MVN with m=2 dimensions (x,y) so the Wishart and Inverse Wishart will have 2m=4 degrees of freedom, so $a_\Sigma, a_3, a_4 = 4$. I will use the data to help with the priors on the hyperparameters. V(x) and V(y) can be equal to the variances of the data or they can be the $\left(\frac{Range(x)}{2}\right)^2$.

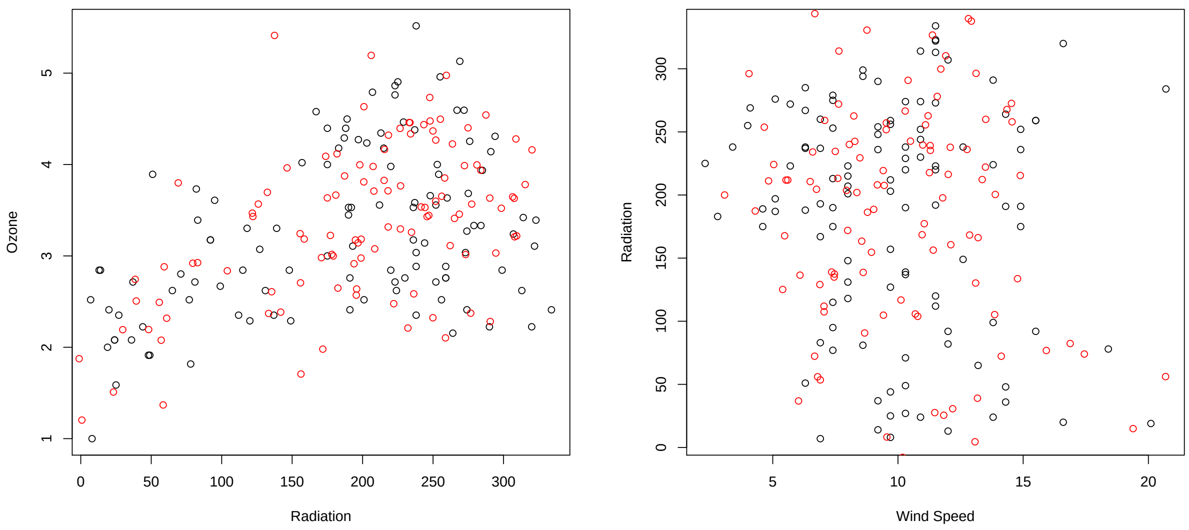
$$\begin{aligned}
(E(x), E(y)) &= \mathbf{a}_1 && (38)\\
(V(x), V(y))I_2 &= E(V(\mathbf{z}|\mu, \Sigma)) + V(E(\mathbf{z}|\mu, \Sigma)) && (39)\\
&= E(\Sigma) + V(\mu) && (40)\\
&= E(E(\Sigma|B_\Sigma)) + V(E(\mu|\mathbf{a}_\mu, B_\mu)) + E(V(\mu|\mathbf{a}_\mu, B_\mu)) && (41)\\
&= E((a_\Sigma - 2 - 1)^{-1}B_\Sigma) + V(\mathbf{a}_\mu) + E(B_\mu) && (42)\\
&= (a_\Sigma - 2 - 1)^{-1}a_4 B_4 + B_1 + (a_3 - 2 - 1)^{-1}B_3 && (43)\\
&= 4B_4 + B_1 + B_3 && (44)
\end{aligned}$$

The parameters having to do with the means will be set to the sample mean of the data. So, $\mathbf{a}_1$ will be set to the data mean. V(x) and V(y) will be placed in a diagonal 2x2 matrix called V_{prior}; we can either use this matrix directly or can multiply it by 5 or 10 to have a more dispersed prior. Overall, $B_4 = \frac{4}{6}V_{prior}$ and $B_1 = \frac{1}{6}V_{prior}$ and $B_3 = \frac{1}{6}V_{prior}$. V_{prior} can also be a diagonal matrix based on the range of the data.

6 Posterior Predictive

We will need L_0 to draw a new y_0:
$p((x_0, y_0)|data) = \int\int MVN((x_0, y_0)|Z(L_0), ...)(\sum_{l=1}^{N} p_l \delta_l(L_0))p(\mathbf{Z}, \mathbf{p}, \mathbf{L}, \alpha, \psi|data)dL_0 d\mathbf{Z} d\mathbf{L} d\mathbf{p} d\psi d\alpha$
Draw L_0 from 1,...,N with probability, $(p_1, p_2, ..., p_N)$ and draw $(x_0, y_0)|L_0 \sim MVN(Z_{l,(1,2)}(L_0), Z_{l,3}(L_0))$

This will provide a set of (x_0, y_0) the same size as the data for every iteration. The following figure is of one set of n=111 posterior predictive observations where the black points are the real data and the red ones are the posterior predictive observations.

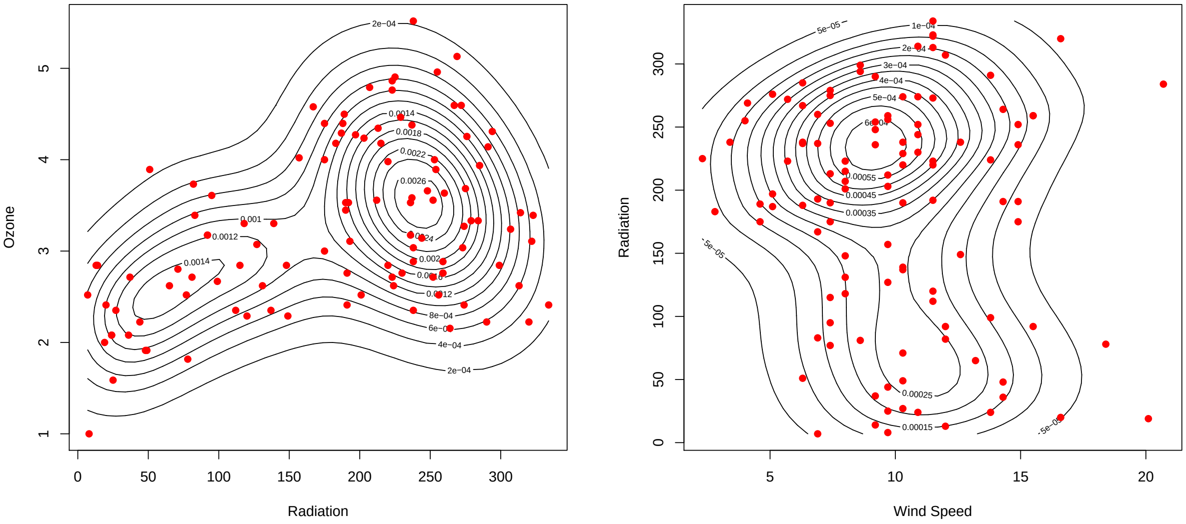


(a) Radiation vs. Ozone

(b) Wind vs Radiation

Figure 2: Posterior Predictive

Then we will look at posterior predictives over a grid of values. This grid will be size 50 by 50. The density at a given point of the grid is: $\sum_{l=1}^{N} p_l * MVN(x_0, y_0|\mu_l, \Sigma_l)$
We will show a contour plot the mean density for all of the iterations. Conditional plots for fixed values of x_0 can be seen as well using these densities.

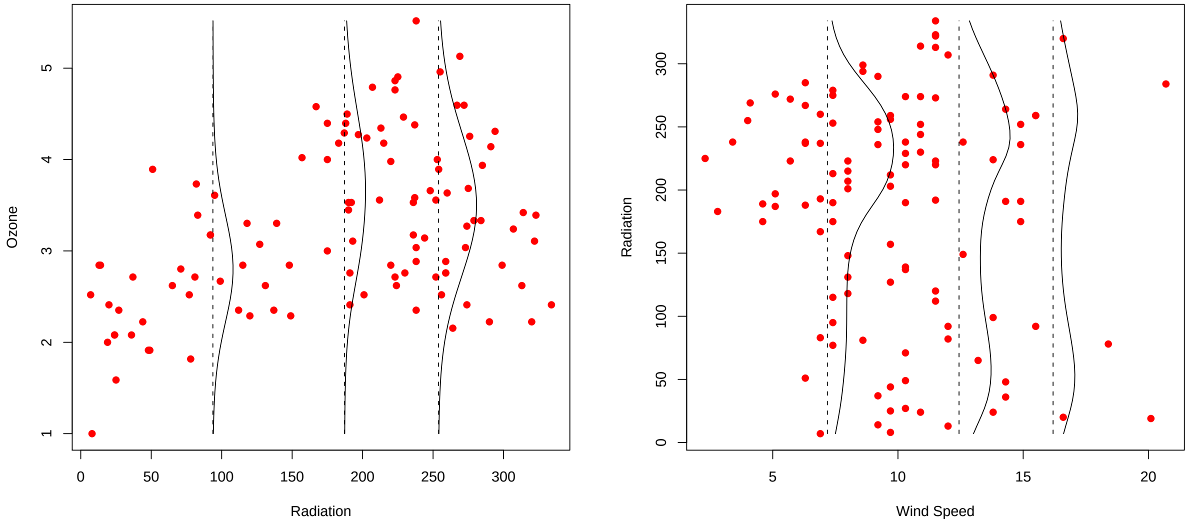


(a) Radiation vs. Ozone

(b) Wind vs Radiation

Figure 3: Posterior Predictive Density

We also want to examine some of the conditionals for some values of x. We can see clearly signs of multimodality in the plot b). And we can see how this changes with different values of x.



(a) Radiation vs. Ozone

(b) Wind vs Radiation

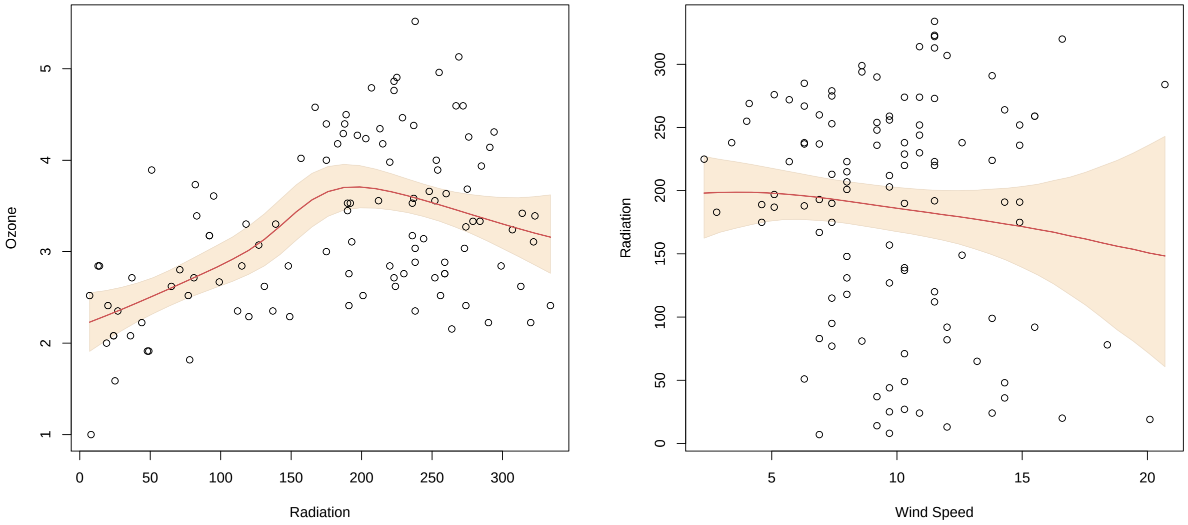
Figure 4: Conditional Posterior Predictive Densities

7 Regression Curve

Lastly, we will observe what a regression type line, $E(Y|X)$, through the data would look like along with 95% probability intervals. The $E(y|x)$ curve leads to a nice calculation of a mixture of Normals with weights based on the stick breaking construction probability and the conditional mean.

$$\begin{aligned}
E(y|x;G_N) &= \int y f(y|x_0;G_N)dy \tag{45}\\
&= \int y \frac{f(y,x_0;G_N)}{f(x_0;G_N)}dy \tag{46}\\
&= \frac{1}{f(x_0;G_N)} \int y \sum_{l=1}^{N} p_l N(y,x_0;G_N)dy \tag{47}\\
&= \frac{1}{f(x_0;G_N)} \int y \sum_{l=1}^{N} p_l N(y|x_0)N(x_0)dy \tag{48}\\
&= \frac{1}{\sum_{l=1}^{N} p_l N(x_0|\mu_l^x,\Sigma_l^x)} \sum_{l=1}^{N} p_l N(x_0|\mu_l^x,\Sigma_l^x) \int yN(y|x_0)dy \tag{49}\\
&= \frac{1}{\sum_{l=1}^{N} p_l N(x_0|\mu_l^x,\Sigma_l^x)} \sum_{l=1}^{N} p_l N(x_0|\mu_l^x,\Sigma_l^x) E(y|x_0) \tag{50}\\
&= \frac{\sum_{l=1}^{N} p_l N(x_0|\mu_l^x,\Sigma_l^x)(\mu_l^y + \Sigma_l^{yx}(\Sigma_l^x)^{-1}(x_0-\mu_l^x))}{\sum_{l=1}^{N} p_l N(x_0|\mu_l^x,\Sigma_l^x)} \tag{51}
\end{aligned}$$

These figures reflect 9000 runs on a 20 x 20 grid. Plot a) is a nice example of a non-parametric fit. Plot b) is not as nice of an example but completely expected since there is no really relationship in the direction we are plotting. A regression curve would make more sense with Radiation as the covariance and Wind speed as the y for plot b).



(a) Radiation vs. Ozone

(b) Wind vs Radiation

Figure 5: $E(Y|X)$

8 Discussion

The goal was to perform density estimation using Bayesian non-parametric methods over a grid and also fit a non-parametric regression type line. Overall, this Dirichlet Process model was chosen because the non-parametric approach easily accommodates multimodal densities. This approach results in posterior predictive plots over a grid of values and even the multimodal conditional plots. A regression type line can be plotted with bands and also carries the non-parametric features of being able to flexibly fit the data in a non-linear way. The drawbacks of this approach is that as the posterior predictive grid grows or the number of covariate grows the algorithm quickly becomes slow.

The block Gibbs algorithm was used here because it allows for easy transition to plotting posterior predictives, as G_N is known. The blocked Gibbs sampler has an added benefit of resampling the cluster values at every step; so we did not need to add any extra steps to the algorithm to do this. This set up employs only Gibbs steps and no Metropolis-Hasting steps, and therefore requires little tuning. Overall, this method was chosen because the alternatives were either estimating G_0 by numerical integration or by using a rough estimate of G_0 to make it conjugate to use the typical Gibbs sampler. These estimated approaches suggested by some authors were considered unfavorable. One other method considered was presented in Neal 2000, which marginalizes over G. This is an acceptable method but we did not use it, as it can take a bit more work at the end to get some of the plots that require G. The Blocked Gibbs sampler was rather straight-forward implement for this model and easily produced some nice plots.

Other things to consider when using a non-parametric model is that it may be affected by the amount, clustering or even the spread of the data. For non-parametric density estimation it is especially important to have enough data. The priors, also, play a role in how the data is fit and how much the prior will influence the results. We used a hierarchical model so there are priors on the cluster mean and variance leading to the updating of these parameters to be more flexible. The amount of dispersion of the prior can effect how the model is able to capture the data and we did not want a prior that was too tight in the center of the data. We made sure to use rather disperse priors to allow the model some flexibility in fitting the data; this included multiplying the variance priors by a constant of 5 or 10. This helped ensure that we did not have a single hump prior right in the center of the data.

As noted, some thought and care should be taken when setting N and also when putting a prior on α; it drives n^*. In our model, we can visually see that there will only be a few clusters maybe 2 or 3; there is an adequate amount of data for so few clusters, n=111. We will most likely have n^* rather small and we know α will be rather small, as well. But by putting a prior on α instead of just setting it at a single value, we allow it to have flexibility to explore the number of clusters that may be in the data. It will be noted that we are not forgetting α's other role in smoothing G. Lastly, the choice of N was considered based on its relationship with α. We assume α will be rather small and we use the typical formula $\epsilon = (\alpha/(\alpha+1))^N$. In practice, we tried several values (10,20,30, and 40) but because the actually number of clusters is somewhere near two and thus α is also small, all of these N values performed in a similarly acceptable fashion; this need not be true with any other data set, especially one with more multimodality.

9 Conclusion

This idea of density estimation and curve fitting is by no means limited to this simplistic example with one covariate. The model can easily be extended to have multiple covariates. Other similar models have been used to model continuous and discrete covariates, binary or multinomial responses, or to do other interesting things like quantile regression. This modeling approach is flexible and can fit a wide range of data types. The implementation of the Blocked Gibbs sampler also has its benefits for these types of models. Here in our simple example we can see some of the posterior predictive density plots and regression line with interval estimates for a one continuous covariate and one continuous response. This model and method seemed to work well for our multimodal density case and provides a flexible Bayesian non-parametric solution to our problem.

10 References

Muller, P., Erkanli, A., and West, M. Bayesian Curve Fitting Using Multivariate Normal Mixtures. Biometrika(1996), 83,1, pp.67-79.

Taddy, M. and Kottas, A. A Bayesian Nonparametric Approach to Inference for Quantile Regression. Technical Report: Department of Applied Mathematics and Statistics, University of California, Santa Cruz, CA.