# Bayesian Nonparametric Regression Models for Modeling and Predicting Healthcare Claims

Robert Richardson

Department of Statistics, Brigham Young University

Brian Hartman

Department of Statistics, Brigham Young University

November 3, 2017

**Abstract**

Standard regression models are often insufficient to describe the complex relationships that are used to predict healthcare claims. A Bayesian nonparametric regression approach is presented as a flexible regression model that relaxes the assumption of Gaussianity. The details for implementation are presented. Bayesian nonparametric regression is applied to a data set of claims by episode treatment group (ETG) with a specific focus on prediction of new observations. It is shown that the predictive accuracy improves compared to standard linear model assumptions. By studying Conjunctivitis and Lung Transplants specifically, it is shown that this approach can handle complex characteristics of the regression error distribution such as skewness, thick tails, outliers, and bimodality.

JEL Codes: C11; C46; I11

**Keywords:** dependent Dirichlet process; episode treatment group; Markov chain Monte Carlo; model comparison, linear models

# 1 Introduction

A number of data-driven problems in insurance can be modeled using regression techniques. These will use a number of covariates, such as age or gender, that relate to an independent variable, such as claim costs or premium rates. A standard regression model, however, inherently assumes characteristics of the data, including independence, Gaussianity, and linearity (Neter et al., 1996). As these assumptions are typically not met, other models have been proposed to do such things as account for outliers (Rousseeuw et al., 1984) and thick tails (Shi, 2013).

The inherent non-Gaussian nature in insurance data has been addressed by using alternative distributions in a generalized linear model such as the gamma and inverse Gaussian (De Jong et al., 2008), the generalized beta (Frees and Valdez, 2008), and others. Other flexible approaches have been proposed such as tweedie regression, quantile regression (Kudryavtsev, 2009), spliced distributions (Gan and Valdez, 2017), and mixture models (Miljkovic and Grün, 2016).

Nonparametric Bayesian modeling has recently been introduced to the actuarial literature as a powerful tool for modeling non-Gaussian densities (Fellingham et al., 2015; Hong and Martin, 2017). This work has shown how powerful the Bayesian nonparametric framework is for handling characteristics of the data such as heavy tails, skewness, or even bimodal distributions. They have also shown an increase in predictive power.

The purpose of this paper is to explore the application of Bayesian nonparametric regression to healthcare claims. The framework that is presented extends past density estimation into flexible distributional assumptions on regression relationships. Bayesian nonparametric regression was introduced in the 90's (Müller et al., 1996). It has since been extended to general ANOVA models (De Iorio et al., 2004) and survival relationships (De Iorio et al., 2009).

This paper will serve to describe methodology for a certain nonparametric Bayesian model which has the potential to improve a number of regression relationships in actuarial applications. The specific utility this regression will also be shown by analyzing a data set of health care costs by episode treatment group. These have been shown to exhibit non-Gaussian densities (Huang et al., 2017). In some cases, adding covariate information accounts for a non-Gaussian density, but as we will see, by adding covariates of age and gender, the error distribution is still considerably non-Gaussian.

While the implementation of Bayesian nonparametric regression presented here will allow the readers to design and use their own algorithms, the DPpackage in R (Jara et al., 2011) already contains a version of Bayesian nonparametric regression that can be used without the need to write up personalized algorithms.

Section 2 provides details for the dependent Dirichlet process ANOVA model used for bayesian nonparametric regression. Details on using this model for a particular analysis given a data set are given in Section 3. The ETG data analysis is shown in Section 4.

# 2 Bayesian Nonparametric Regression

A simple regression model with a single covariate could be written as

$$y_i = f(x_i) + \epsilon.$$

A number of methods of fitting this model could be considered Bayesian nonparametric regression. For example, $f(x_i)$ could have a parametric structure and $\epsilon$ has a DP prior, such that $\epsilon|G \sim G$, and $G \sim DP$. On the other hand, $\epsilon$ could be a standard parametric distribution, such as $N(0, \sigma^2)$, and $f(x_i)$ could have a flexible mean structure, such as using a basis function expansion or splines (Eilers and Marx, 1996; Vidakovic, 1998). Such procedures as Gaussian process regression (Gramacy and Lee, 2008) and regression trees (Chipman et al., 1998) are also considered semiparametric or nonparametric.

Here we consider fully nonparametric regression from a Bayesian standpoint. The general idea is to apply a dependent Dirichlet process (DDP) (MacEachern, 1999) to the joint parameter space of the regression coefficients. In this approach, the standard parametric forms of both the mean function and the error process are replaced with a more flexible structure.

## 2.1 Dependent Dirichlet Process

A dependent Dirichlet process is the basis for fully nonparametric regression. DDPs can be considered a prior on families of random probability measures on some domain

$D$. Let $G_D$ be a DDP$(\alpha, G_{0,D})$. Then

$$G_D = \sum_{l=1}^{\infty} w_l \delta_{\theta_{l,D}}$$

where each $\theta_{l,D} = \{\theta_l(x) : x \in D\}$ are independent realizations from a stochastic process $G_{0,D}$ which lives on the domain $D$ and the weights arise from stick-breaking where $w_l = \xi_l \prod_{i=1}^{l-1}(1 - \xi_i)$ and $\xi \overset{i.i,d,}{\sim} Beta(1, \alpha)$. The main difference between DDPs and standard Dirichlet processes is that the point masses are actually realizations from a base stochastic process, $G_{0,D}$, as opposed to some base distribution.

The distribution of a finite number of points $\mathbf{x} = (x_1, x_2, ..., x_n)$ in the domain $D$ can be constructed as a mixture of draws from the finite dimensional distribution of $G_{0,D}$, meaning that if $f(x) = G_D$ then $(f(x_1), f(x_2), ...., f(x_n)) \sim G_{\mathbf{x}}$ where $G_{\mathbf{x}} = \sum_{l=1}^{\infty} w_l \theta_{l,\mathbf{x}}$ where $\theta_{l,\mathbf{x}} \overset{i.i.d.}{\sim} G_{0,\mathbf{x}}$ and $G_{0,\mathbf{x}}$ is a multivariate distribution arising from the joint distribution of finite points on $G_{0,D}$. A typical example of this is when $G_{0,D}$ is a Gaussian process. Then $G_{0,\mathbf{x}}$ is a multivariate normal distribution with mean and variance as functions of the points in $\mathbf{x}$ according to the mean function and covariance structure of $G_{0,D}$.

Like standard Dirichlet processes, DDPs can be mixed with distributions for continuous covariates. Consider a mixture of multivariate normal distributions where the mean vector is mixed with a DDP prior. Then if $\mathbf{y} = (f(x_1), f(x_2), ..., f(x_n))$ then

$$\begin{aligned}\mathbf{y} &\sim g(\mathbf{y}),\\ g(y) &= \int \phi(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) dG_{\mathbf{x}}(\boldsymbol{\mu}),\end{aligned}$$

where $\boldsymbol{\mu} = (\mu(x_1), \mu(x_2), ..., \mu(x_n))$ and $G_{\mathbf{x}}(\boldsymbol{\mu}) = \sum_{i=1} w_l \delta_{\boldsymbol{\theta}_{l,\mathbf{x}}}$ where $\boldsymbol{\theta}_{l,\mathbf{x}} \overset{i.i.d.}{\sim} G_{0,\mathbf{x}}$ are multivariate realizations from a base joint distribution $G_{0,\mathbf{x}}$. A practical way of writing this is

$$\begin{aligned}\mathbf{y} &\sim g(\mathbf{y}),\\ g(y) &= \sum_{l=1}^{\infty} w_l \phi(\mathbf{y}|\boldsymbol{\mu}_l, \boldsymbol{\Sigma})\\ \boldsymbol{\mu}_l &\overset{i.i.d}{\sim} G_{0,\mathbf{x}}, \quad l = 1, 2, ..., .\end{aligned}$$

As in all these examples, the weights, $w_l$ arise from stick-breaking. Another extension is to include the covariance matrix $\boldsymbol{\Sigma}$ as atoms in the DDP, leading to both $\boldsymbol{\mu}_l$ and $\boldsymbol{\Sigma}_l$ being drawn jointly from the base distribution $G_{0,\mathbf{x}}$.

## 2.2 DDP ANOVA

These ideas can be extended to a regression setting. A DDP ANOVA extends DDPs to include covariate information (De Iorio et al., 2004, 2009). Let $\mathbf{z}_i$ be a vector of covariate information for a specific record, $\mathbf{z}_i = (1, z_{i,1}, z_{i,2}, ..., z_{i,p})'$. Then a DDP ANOVA model for $\mathbf{y} = (y_1, ..., y_n)'$ is

$$\begin{aligned}\mathbf{y} &\sim g(\mathbf{y}) && (1)\\ g(\mathbf{y}) &= \sum_{l=1}^{\infty} w_l \phi(\mathbf{y}|\mathbf{z}'\boldsymbol{\beta}_l, \boldsymbol{\Sigma}_l) && (2)\\ (\boldsymbol{\beta}_l, \boldsymbol{\Sigma}_l) &\sim G_0(\psi) \quad l = 1, 2....., \quad \psi \sim \pi(\psi) && (3)\\ w_l &= \xi_l \prod_{i=1}^{l-1}(1-\xi_i), \quad \xi \overset{i.i.d,}{\sim} Beta(1, \alpha), \quad l = 1, 2, ... && (4)\end{aligned}$$

In this model, the atoms drawn from the base distribution are the regression coefficients and the covariance. A few common simplifications are setting $\boldsymbol{\Sigma}_l = \sigma_l^2 \mathbf{I}_n$ and then constructing $G_0$ such that $\sigma_l^2$ and $\boldsymbol{\beta}_l$ are *a priori* independent. The base distribution will also have parameters $\psi$ that can have hyperprior, $\pi(\psi)$.

The result of this construction of a regression model is flexible relationships between the covariates and the independent variable and a flexible error structure. Conditional on a certain atom index $l$, the mode is a normal linear regression model, but by mixing on the infinite set of all atoms, the normal mixture has full support on the entire space of covariate and error distributions, which essentially means that there is no regression relationship that the DDP ANOVA model cannot represent.

For this paper we use the following base distribution and hyperpriors.

$$\begin{aligned} G_0 &= N(\boldsymbol{\beta}|\mu_\beta, \Sigma_\beta) \times \mathrm{IG}(\sigma^2|a_\sigma, b_\sigma) \\ \mu_\beta, \Sigma_\beta &\sim \mathrm{NIW}(\mu_\beta, \Sigma_\beta|\mu_0, \kappa_0, \nu_0, \Psi_0) \\ a_\sigma &\sim \mathrm{Gamma}(a_\sigma|\zeta_a, \eta_a), \quad b_\sigma \sim \mathrm{Gamma}(b_\sigma|\zeta_b, \eta_b), \end{aligned}$$

Where IG represent an inverse gamma distribution and NIW represent a normal-inverse Wishart distribution. Hyperprior values must be set for $\mu_0, \kappa_0, \nu_0, \Psi_0, \zeta_a, \eta_a, \zeta_b$, and $\eta_b$. The parameter $\alpha$ can be learned using a prior or simply fixed. Depending on the size of the data, the choices for hyperpriors may play an important role in the he results of the analysis, so they must be chosen carefully. $\mu_0$ and $\Psi_0$ must be chosen to represent prior belief in the regression coefficients and the covariance, where $\kappa_0$ and $\nu_0$ are chosen to represent the respective confidence in the prior belief of $\mu_0$ and $\Psi_0$. The gamma hyperpriors could be chosen to yield expected values for $a_\sigma$ and $b_\sigma$ that represents belief in the variance of the regression model.

To aid in choosing priors that fit with interpretations, the data should be standardized, meaning that the mean of variable is subtracted off of each data point in that variable and is divided by the standard deviation of the variable. Results will not be affected by this transformation as they can easily be transformed back to the original scale to make any specific inference.

## 3 Posterior Inference

For a regression problem with $n$ observations and a $p$ predictor variables, a fully non-parametric approach to Bayesian regression can be achieved through the model in equations (1) through (4). The key features of the estimation procedure is a Gibbs sampler where each unknown variable is drawn from conditional distributions given all the other parameters.

The infinite sum in Equation (2) is approximated by a finite sum, which can only be done with careful consideration of the expected number of components. This allows the individual data points to be assigned to specific latent clusters, providing nearly all parameter updates to be conjugate. Details are found in Appendix A and the steps of prediction of new observations, which is used extensively in the next section, is found in Appendix B.

## 4 Modeling Claims by Episode Treatment Group

Episode Treatment Groups (ETG) is a classification scheme for a variety of conditions that require medical services. ETGs are used by to help predict the future costs of a particular book of business.

As health insurers are also interested in the uncertainty associated with predictions of future costs, an accurate representation of the distributional characteristics of the ETG summaries is important. This idea is explored for ETGs in Huang et al. (2017) where a number of different modeling approaches were taken to model the ETG densities. We extend that exploration here using Bayesian nonparametric regression by adding covariate information and estimating regression relationships as opposed to densities. This approach is shown here as an illustration of the usefulness of Bayesian nonparametric regression and not necessarily a case study of ETG behavior.

Each record has two covariates, age and health, along with the healthcare charges. Age will be treated as a continuous variable. The summary statistics of these covariates vary widely based on the ETG as some diseases mainly impact certain demographic, such as pregnancy. In the cases where gender is incredibly skewed in favor of one gender or another, gender was left out as a covariate and only age was used.

To display the breadth of the advantage Bayesian nonparametric regression affords, we will analyze the data for all 347 ETGs. However, because the data sets can be quite large, only subsets are used in most cases. Both Bayesian nonparametric regression and a Bayesian linear model are fit for each ETG. We then explore the attributes of the models with two specific ETGs: Conjunctivitis and Lung Transplants, a large and small sample respectively.

## 4.1 Results for Subsets of 347 ETGs

For each ETG, a subset of 1,100 data points were chosen at random, or all the data points were used if the size of the group was less than 1,000. The DDP ANOVA model as well as a Bayesian linear regression model was fit to 1,000 data points of each subset and then 100 was left out to explore the predictive accuracy of the models. The prior values for the DDP ANOVA model are $\mu_0 = (0, 0, 0)'$, $\Psi_0 = \mathbf{I}_3, \kappa_0 = 10, \nu_0 = 10, \zeta_a = 10, \eta_a = 2, \zeta_b = 10$, and $\eta_b = 2$. These priors were chosen to match the standardized data, so without covariate information we expect the data to have mean 0 and variance 1, and also to reflect the proper uncertainty in the parameters. The posterior draws were only sensitive to prior information in the cases where the sample size of data was smaller than about 100.

The Bayesian linear model was also fit. This model was constructed in a similar way as the DDP ANOVA model would be if there were one cluster with all the data. As such, we use all the same priors and hyperpriors for the parameters $\boldsymbol{\beta}$ and $\sigma_2$.

Three metrics are used to determine the quality of the model fit. The first is the mean squared prediction error (MSPE) for the observations that were held back from the model fit. While MSPE is an important tool to evaluate prediction, it only accounts for the mean of the posterior predictive distribution. To more accurately asses the overall quality of the posterior predictive samples we also use continuous rank probability score (CRPS) to evaluate prediction accuracy. If $y^{(1)}, ...., y^{(B)}$ are $B$ samples from a fitted distribution and $y_T$ is an observation, then CRPS is

$$CRPS_y = \frac{1}{n}\sum_{i=1}^{B}(y^{(i)} - y_T)^2 + \frac{1}{n^2}\sum_{i=1}^{B}\sum_{j=1}^{B}(y^{(i)} - y^{(j)})^2 \tag{5}$$

The last metric used is to assess penalized in-sample model fit. This is the Deviance Information Criterion. It is analogous to Akeike's information criterion and Bayesian information criterion for model fits, but is more appropriate for Bayesian output when the model fit is given in terms of samples from the posterior distribution of parameters. For all three of these metrics, lower values are better.

Figure 1 shows the MSPE and the CRPS for every ETG used in the analysis. The MSPE is nearly indistinguishable between the two models, meaning that electing to have more flexible distributional assumptions did not improve the point estimate much. However, the CRPS, which takes into account the distributional fit, suggests that the Bayesian non-parametric regression significantly improves the prediction performance.

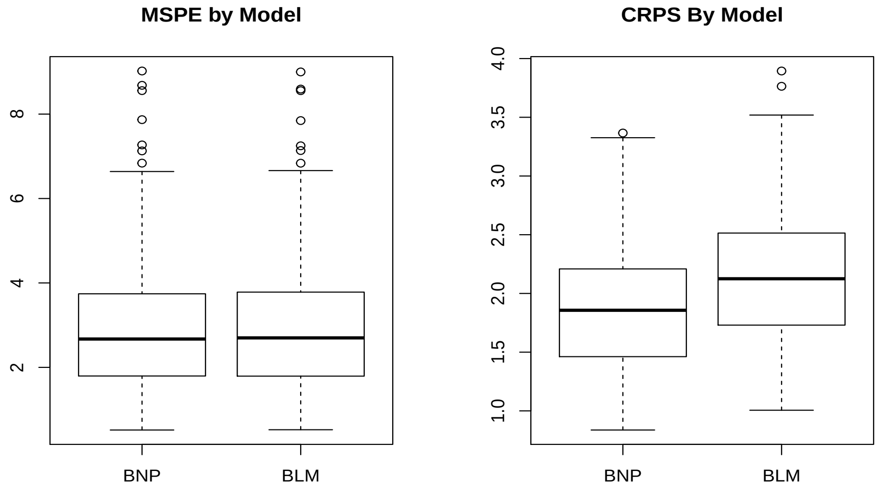

Figure 1: CRPS and MSPE values are shown for the BNP regression model and for the Bayesian linear model. The CRPS shows a significant improvement using the BNP and the MSPE shows little to no improvement.

| Metric | BNP | BLM |
|---|---|---|
| Average MSPE | 3.002 | 3.013 |
| Average CRPS | 1.868 | 2.146 |
| Count Lower MSPE | 161 | 159 |
| Count Lower CRPS | 316 | 4 |

Table 1: Some key comparisons between the prediction metrics for the two models.

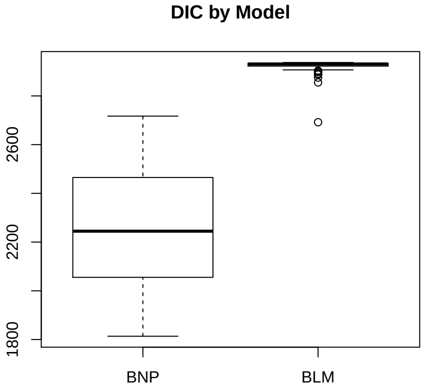

Figure 2: The DIC value is shown for the two models. The BNP regression model is consistently better.

Table 1 shows some additional values to support these conclusions. On average, the MSPE for the BNP model was only 2% better, but the CRPS for the BNP model was 14% better on average than the standard linear model. The implication of this result is that while point estimates of future predictions may not suffer as much from poor distributional assumptions, results that rely on the predictive distribution, such as percentiles, confidence intervals, and a variety of other things that healthcare insurers are interested in, will be largely affected.

For comparing the in-model fits, the DIC favored the BNP regression model for every single ETG, suggesting that the extra parameters required in the BNP regression model were greatly improving the model fit in every case. The average improvement for the DIC from using BNP regression was 38%. While prediction was clearly affected by the more flexible model fit, in-sample diagnostics show an even greater improvement.

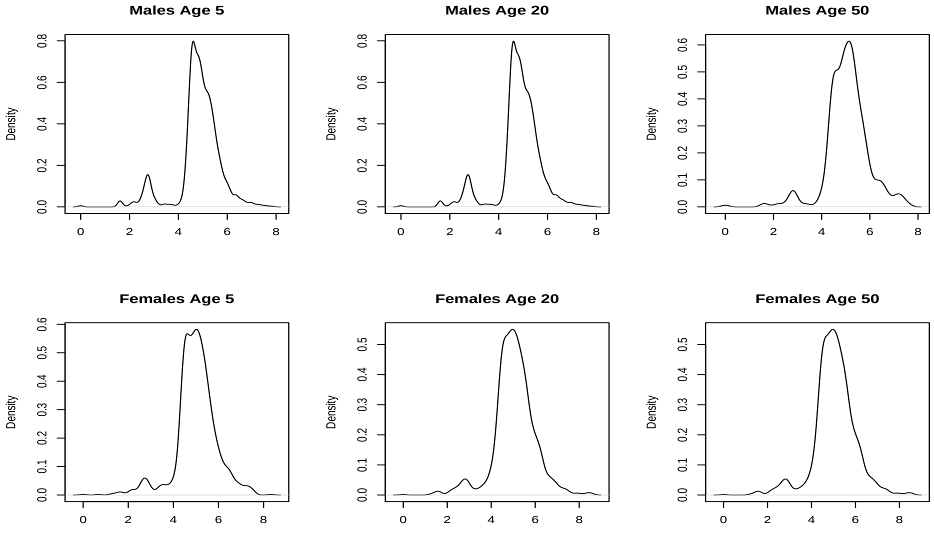

Figure 3: Empirical densities of the total charges for the conjunctivitis ETG for a number of covariate combinations.

## 4.2 Conjuctivitis

Conjunctivitis was chosen as a special case to explore because it has many observations and the distributional features are especially non-Gaussian. The distribution is even bimodal in many cases. This non-Gaussian behavior can be seen clearly for the log charges in Figure 3. The left tail is wider than a Gaussian tail and there is an extra mode. With the information included in the study, it seems that this mode cannot be explained by the covariates alone and a more flexible distributional assumption is appropriate.

A Bayesian linear model and the DDP ANOVA model were fit to the training data set, which comprised 90% of the data, a total of 160,228. The other 10% was left out to compute MSPE and CRPS to analyze predictive accuracy. The median values and 95% credible intervals for the coefficients using the Bayesian linear model are shown in Table 2 along with the average effects from the nonparametric regression model, where

| | | 2.5% | 50% | 97.5% |
|---|---|---|---|---|
| BLM | $\hat{\beta}_1$ | 0.0791 | 0.0841 | 0.0890 |
| | $\hat{\beta}_2$ | -0.01 | $-0.000008$ | 0.01 |
| BNP | $\hat{\beta}_1$ | 0.0714 | 0.0768 | 0.820 |
| | $\hat{\beta}_2$ | 0.0032 | 0.011 | 0.0187 |

Table 2: Median and 95% credible interval for the sampled coefficients for the BLM and the average effects of the BNP regression model.

$\hat{\beta}_1$ is the coefficient for age and $\hat{\beta}_2$ is the coefficient for gender. The average effects for the BNP model is found for each sample by taking a weighted average of the atoms, $\sum_{l=1}^{N} w_l \boldsymbol{\beta}_l$. Table 2 shows that the Gaussian assumption leads to no effect of gender, where the BNP model is able to detect an effect, although it is small.

Table 3 lists the comparison metrics for the two model fits. Even with the bimodal error structure, the point estimate for the linear model is not far off from the more flexible model. The CRPS and DIC are both considerably better for the BNP regression model. The posterior predictive distribution is plotted for both models for an 18 year old female in Figure 4. The posterior predictive distributions are overlaid the histogram of all 18 year females in the data set. The extra bump in the left tail is accurately captured by the posterior predictive of the BNP regression model. To try and capture the tail, the BLM model gives a wider prediction than is necessary.

| Metric | BNP | BLM |
|---|---|---|
| MSPE | 0.8375 | 0.8371 |
| CRPS | 0.938 | 0.998 |
| DIC | -398,000 | -250,000 |

Table 3: Some key comparisons between the prediction metrics for the two models for the Conjunctivitis ETG.

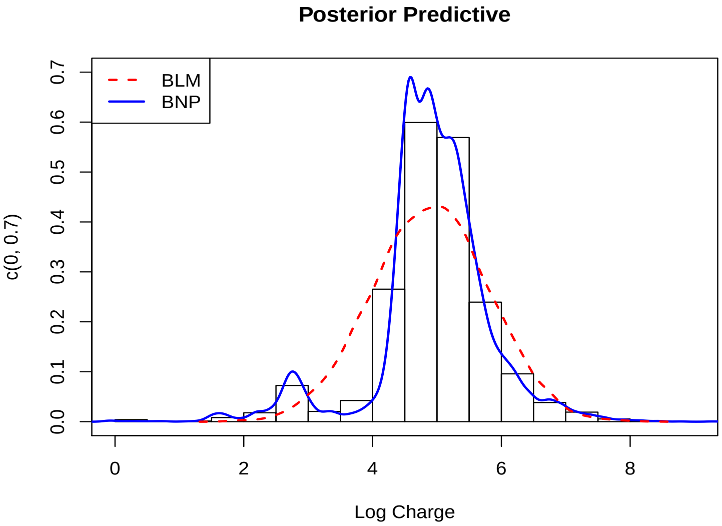

Figure 4: The posterior predictive distribution for the log charges of an 18 year old female with conjunctivitis overlaid the histogram of the data of all 18 year females in the ETG data set.

## 4.3 Lung Transplant

Lung transplant was also chosen to examine more carefully because it is a smaller data set. The data cannot be subsetted the same way to find complete histograms of certain covariate combinations, as was the case for conjunctivitis. The data is still skewed, which means that it is unlikely that the Gaussian assumption for the error structure is appropriate. To account for the outliers, the BNP model will predict thicker tails in the posterior predictive distribution. The BLM model will just get a wider variance.

Again, several data points were left out of the analysis to be used in prediction. The posterior predictive distribution for 4 of those individuals are shown in Figure 5. The wide variance in the BLM predictions can be seen. The thicker tails in the BNP regression model is difficult to detect by eye, but they are thicker tails than a standard Gaussian. In all 4 cases, the actual observation, which, again, wasn't used in the model

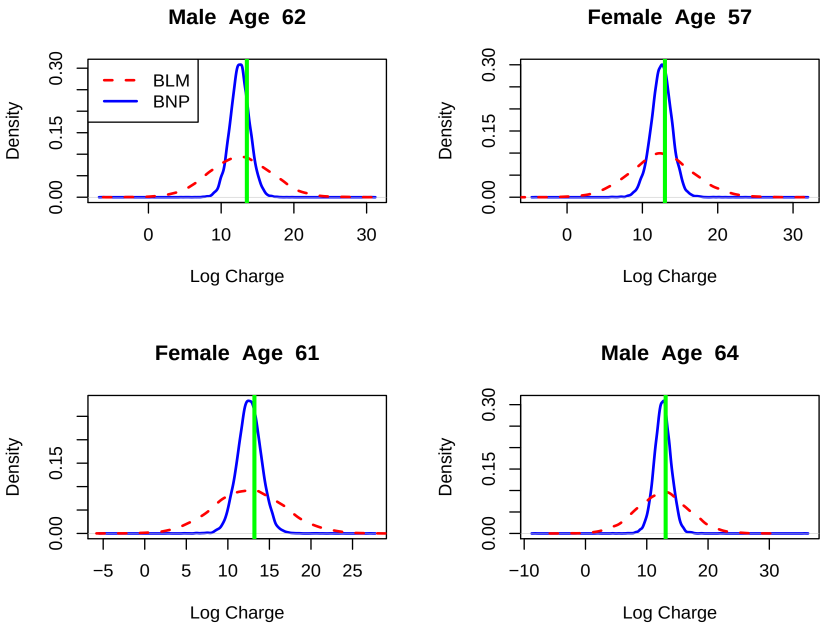

Figure 5: The posterior predictive distribution for the log charges of 4 individuals left out of the analysis for prediction for the Lung Transplant ETG.

fit, is well within the bounds of both prediction intervals. This actual observation is shown in the plots by a vertical spike.

| Metric | BNP | BLM |
|---|---|---|
| MSPE | 0.535 | 0.524 |
| CRPS | 1.246 | 3.426 |
| DIC | 98 | 172 |

Table 4: Some key comparisons between the prediction metrics for the two models for the Lung Transplant ETG.

The MSPE is actually lower for the BLM model, as seen in Table 4. But the wider variance leads to poor distributional accuracy and a much higher CRPS and DIC for the predictions.

# 5 Conclusion

The utility of a non-Gaussian regression relationship has been illustrated for healthcare data in a number of past work. This paper has served to introduce Bayesian nonparametric regression as a very powerful tool for flexible regression. It can be used to model a wide variety of error terms. The DDP ANOVA model is useful for continuous variables as the independent variable. Other nonparametric Bayesian models have been introduced for other variable types including in multivariate and mixed-type settings (Kottas et al., 2005; Dunson and Xing, 2009; DeYoreo et al., 2015).

Other extensions could be applied from the literature. One in particular that may be useful is to make the weights of the DDP be dependent on the covariate information. The mixture used in the error distribution can then be covariate dependent. For example, in the conjunctivitis example,the younger patients had a stronger mode in the left tail than the older patients. This can be modeled explicitly by making the weights depend on the covariates.

As mentioned in the introduction, a version of Bayesian nonparametric regression is contained within the DPpackage in R (Jara et al., 2011). This allows these techniques to be used without the effort of constructing personalized algorithms.

# A Blocked Gibbs Sampler

As with several problems in Bayesian statistics, inference for the DDP ANOVA model is done by generating posterior samples of the unknown parameters. There are a variety of methods of sampling from the posterior distribution of atoms from a dependent Dirichlet process model. The one we will use here is called blocked Gibbs sampling. The main benefit of this method is computational simplicity. Theoretically, the other sampling techniques such as using full conditionals or slice sampling will yield similar results.

## A.1 Finite Approximation

When using blocked Gibbs sampling the number of atoms in infinite mixture seen in equation (2) is truncated to $N$ distinct components, where $N < n$. The danger in this is that if $N$ is chosen to be too low, then there will not be enough predetermined clusters as s needed for the data. For a specific value of $\alpha$ the approximate expected value of $N$ is $E(N|\alpha) \approx \alpha \log\left(\frac{\alpha+n}{\alpha}\right)$ and the approximate variance is $Var(N|\alpha) \approx \alpha\left(\log\left(\frac{\alpha+n}{\alpha}\right) - 1\right)$. This information could be useful to determine an appropriate value to fix $N$. For example if there are 1000 data points and $\alpha = 3$, then the expected value for $N$ is approximately 17.4 and two standard deviations above that is approximately 25, so setting $N$ to a number larger than 25 would be reasonable. In the posterior samples it will be possible to check the number of clusters that were actually used. If that number is close to or equals $N$ in some of the samples, the value for $N$ may not have been adequate and the analysis should be redone with a larger number of fixed clusters.

With the truncation of clusters, there will now be only $N$ weights, with the requirement that $\sum_{l=1}^{N} w_l = 1$. To ensure this, only the first $N - 1$ weights will be found through stick-breaking. The final one will be set to be the remainder, $w_N = 1 - \sum_{i=1}^{N-1} w_l = \prod_{l=1}^{N-1}(1 - \xi_l)$. The expected value of this final weight will be $E(w_N|\alpha) = (\alpha/(\alpha+1))^{N-1}$. So as $N$ increases, this value will get closer to 0, which is desirable for consistency in the number of clusters for our choice of $N$. As $\alpha$ increases, this value goes up, which means a higher $N$ is needed to ensure that this value is close to 0.

## A.2 Latent Assignment Variables

Another feature of blocked Gibbs sampling is latent assignment variables. There is one assignment variable for every observation, $L_1, ..., L_n$. These can take in integer values between 1 and $N$, assigning each observation to one of the $N$ clusters. Equations (1) and (2) in this case can be rewritten as

$$\begin{aligned} y_i|\mathbf{z}_i, L_i &\sim N(y_i; \mathbf{z}_i'\boldsymbol{\beta}_{L_i}, \sigma^2_{L_i}) && (6) \\ L_i|\mathbf{w} &\sim \sum_{i=1}^{N} w_l \delta_l(L_i) && (7) \end{aligned}$$

## A.3 Gibbs Sampling

The actual samples will be taken from the posterior using Gibbs sampling, which is sampling a subset of the variables conditional on the data and the most recent sample of all the other variables and then rotating through other subsets of the variables. The subsets we use are (1) the $N$ atoms of regression coefficients $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_N$, (2) the $N$ variance atoms, $\sigma^2_1, ..., \sigma^2_N$, (3) the weights $w_1, ..., w_N$, (4) the assignment variables $L_1, ..., L_n$, (5) the hyperpriors $\mu_\beta, \Sigma_\beta, a_\sigma$, and $b_\sigma$, and (6) the value for $\alpha$.

1. The regression coefficient atoms will be sampled one at a time. For coefficient $\boldsymbol{\beta}_l$, the data itself is subsetted. Let $\mathbf{y}^{(l)}$ and $\mathbf{z}^{(l)}$ be the subsetted independent variable and design matrix respectively where record $j$ is included only if $L_j = l$. The samples are drawn for $\boldsymbol{\beta}_l$ from the distribution

$$\boldsymbol{\beta}_l|\cdot \sim N(\boldsymbol{\beta}_l|\mu^*_\beta, \Sigma^*_\beta) \tag{8}$$

   where $\Sigma^*_\beta = (\mathbf{z}^{(l)}\mathbf{z}^{(l)} + \Sigma_\beta^{-1})^{-1}$ and $\mu^*_\beta = \Sigma^*_\beta(\mathbf{z}^{(l)\prime}\mathbf{y}^{(l)} + \Sigma_\beta^{-1}\mu_\beta)$.

2. The subsetted data will also be used to draw samples for $\sigma_1^2, ..., \sigma_N^2$, using the conditional posterior

$$\sigma_l^2|\cdot \sim \text{IG}\left(\sigma_l^2|a_\sigma + M_l/2, b_\sigma + .5\left(\sum_{L_i=l} y_i^2 + \mu_\beta'\Sigma_\beta^{-1}\mu_\beta + \mu_\beta'^*\Sigma_\beta^{*-1}\mu_\beta^*\right)\right) \tag{9}$$

   where $M_l = |L_i = l$ is the size of the subset.

3. The weights are found using stick-breaking, although conditional on the alignment variables, $\xi_l \sim \text{Beta}(1 + M_l, \alpha + \sum_{j=l+1}^n M_j)$ for $l = 1, ..., N-1$. Then $w_l = \xi_l \prod_{i=1}^{l-1}(1-\xi_i)$ for $l = 1, ..., N-1$ and $w_N = 1 - \sum_{i=1}^{N-1} w_i$

4. The assignment variables are drawn from a discrete distribution where

$$Pr(L_i = l) \propto w_l \phi(y_i|\mathbf{z}_i'\boldsymbol{\beta}_l, \sigma_l^2).$$

5. The base distribution is assumed to be separate in our formulation although it could easily be whatever the user wishes. For $\mu_\beta$ and $\Sigma_\beta$, posterior samples can be taken from

$$\begin{aligned}\mu_\beta, \Sigma_\beta|\cdot \sim &\text{NIW}(\mu_\beta, \Sigma_\beta|\frac{1}{\kappa_0 + N}(\kappa_0\mu_0 + n\bar{\boldsymbol{\beta}}), \kappa_0 + n, \nu_0 + n,\\ &\psi + \sum_{l=1}^N(\boldsymbol{\beta}_l - \bar{\boldsymbol{\beta}})(\boldsymbol{\beta}_l - \bar{\boldsymbol{\beta}})' + \frac{\kappa_0 N}{\kappa_0 + N}(\bar{\boldsymbol{\beta}} - \mu_0)(\bar{\boldsymbol{\beta}} - \mu_0)')\end{aligned}$$

   The posterior samples for $b_\sigma$ are drawn from

$$b_\sigma|\cdot \sim \text{Gamma}(b_\sigma|\zeta_b + Na_\sigma, \eta + \sum_{l=1}^N 1\sigma_l^2)$$

   There is no conjugate sampler for $a_\sigma$. It can be drawn using a Metropolis-

Hastings algorithm, where a proposal is made for a new value of $a_\sigma$ given the previous value. If this generating distribution is $g(a_\sigma^*|a_\sigma)$ then the new value $a_\sigma^*$ is accepted with probability

$$\min\left(1, \frac{\prod_{l=1}^{N} \text{IG}(\sigma_l^2|a_\sigma^*, b_\sigma)\text{Gamma}(a_\sigma^*|\zeta_a, \eta_a)g(a_\sigma|a_\sigma^*)}{\prod_{l=1}^{N} \text{IG}(\sigma_l^2|a_\sigma, b_\sigma)\text{Gamma}(a_\sigma|\zeta_a, \eta_a)g(a_\sigma^*|a_\sigma)}\right)$$

6. If given a Gamma$(a_\alpha, b_\alpha)$ prior, posterior samples for $\alpha$ can be drawn from

$$\alpha|\cdot \sim \text{Gamma}(N + a_\alpha - 1, b_\alpha - \log(w_N))$$

By repeating steps 1 through 6 iteratively, samples for each of the parameters will be collected.

# B Prediction of New Observations

Frequently of interest in modeling claims data is to be able to predict from the model given a certain set of predictor variables, $\mathbf{z}^*$. If $B$ samples are drawn from the posterior distribution of the parameters then the predictive distribution for a new observation can be found using the following steps for $b = 1, ..., B$, meaning variables superscripted by $(b)$ are the $b$-th sample.

1. Draw a value, $l^*$, between 1 and $N$ with probability $Pr(l^* = j) = w_j^{(b)}$
2. Set $\boldsymbol{\beta}^*$ equal to $\boldsymbol{\beta}_{l^*}$ and set $\sigma^{*2}$ equal to $\sigma_{l^*}^2$
3. Draw a new value $y^{*(b)}$ from $N(\cdot|\mathbf{z}^{*\prime}\boldsymbol{\beta}^*, \sigma^{*2})$

The result is a sample from the posterior predictive distribution given covariates $\mathbf{z}^*$.

# References

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998), "Bayesian CART model search," *Journal of the American Statistical Association*, 93, 935–948.

De Iorio, M., Johnson, W. O., Müller, P., and Rosner, G. L. (2009), "Bayesian nonparametric nonproportional hazards survival modeling," *Biometrics*, 65, 762–771.

De Iorio, M., Müller, P., Rosner, G. L., and MacEachern, S. N. (2004), "An ANOVA model for dependent random measures," *Journal of the American Statistical Association*, 99, 205–215.

De Jong, P., Heller, G. Z., et al. (2008), *Generalized linear models for insurance data*, vol. 10, Cambridge University Press Cambridge.

DeYoreo, M., Kottas, A., et al. (2015), "A fully nonparametric modeling approach to binary regression," *Bayesian Analysis*, 10, 821–847.

Dunson, D. B. and Xing, C. (2009), "Nonparametric Bayes modeling of multivariate categorical data," *Journal of the American Statistical Association*, 104, 1042–1051.

Eilers, P. H. and Marx, B. D. (1996), "Flexible smoothing with B-splines and penalties," *Statistical science*, 89–102.

Fellingham, G. W., Kottas, A., and Hartman, B. M. (2015), "Bayesian nonparametric predictive modeling of group health claims," *Insurance: Mathematics and Economics*, 60, 1–10.

Frees, E. W. and Valdez, E. A. (2008), "Hierarchical insurance claims modeling," *Journal of the American Statistical Association*, 103, 1457–1469.

Gan, G. and Valdez, E. A. (2017), "Fat-Tailed Regression Modeling with Spliced Distributions," , Available at SSRN: https://ssrn.com/abstract=3037062.

Gramacy, R. B. and Lee, H. K. H. (2008), "Bayesian treed Gaussian process models with an application to computer modeling," *Journal of the American Statistical Association*, 103, 1119–1130.

Hong, L. and Martin, R. (2017), "A flexible Bayesian nonparametric model for predicting future insurance claims," *North American Actuarial Journal*, 21, 228–241.

Huang, S., Hartman, B., and Brazauskas, V. (2017), "Model Selection and Averaging of Health Costs in Episode Treatment Groups," *ASTIN Bulletin: The Journal of the IAA*, 47, 153–167.

Jara, A., Hanson, T. E., Quintana, F. A., Müller, P., and Rosner, G. L. (2011), "DPpackage: Bayesian semi-and nonparametric modeling in R," *Journal of statistical software*, 40, 1.

Kottas, A., Müller, P., and Quintana, F. (2005), "Nonparametric Bayesian modeling for multivariate ordinal data," *Journal of Computational and Graphical Statistics*, 14, 610–625.

Kudryavtsev, A. A. (2009), "Using quantile regression for rate-making," *Insurance: Mathematics and Economics*, 45, 296–304.

MacEachern, S. N. (1999), "Dependent nonparametric processes," in *ASA proceedings of the section on Bayesian statistical science*, Alexandria, Virginia. Virginia: American Statistical Association; 1999, pp. 50–55.

Miljkovic, T. and Grün, B. (2016), "Modeling loss data using mixtures of distributions," *Insurance: Mathematics and Economics*, 70, 387–396.

Müller, P., Erkanli, A., and West, M. (1996), "Bayesian curve fitting using multivariate normal mixtures," *Biometrika*, 83, 67–79.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996), *Applied linear statistical models*, vol. 4, Irwin Chicago.

Rousseeuw, P., Daniels, B., and Leroy, A. (1984), "Applying robust regression to insurance," *Insurance: Mathematics and Economics*, 3, 67–72.

Shi, P. (2013), "Fat-tailed regression models," *Predictive Modeling Applications in Actuarial Science*, 1, 236–259.

Vidakovic, B. (1998), "Nonlinear wavelet shrinkage with Bayes rules and Bayes factors," *Journal of the American Statistical Association*, 93, 173–179.