Insurance: Mathematics and Economics 83 (2018) 1–8

Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

journal homepage: www.elsevier.com/locate/ime

Bayesian nonparametric regression models for modeling and predicting healthcare claims

Robert Richardson *, Brian Hartman

Department of Statistics, Brigham Young University, United States

ARTICLE INFO

Article history:
Received November 2017
Received in revised form June 2018
Accepted 12 June 2018
Available online 18 June 2018

JEL classification:
C11
C46
I11

Keywords:
Dependent Dirichlet process
Episode treatment group
Markov chain Monte Carlo
Model comparison
Linear models

ABSTRACT

Standard regression models are often insufficient to describe the complex relationships that exist in 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 dataset of claims by episode treatment group (ETG) with a specific focus on prediction of new observations. It is shown that the predictive accuracy improves when compared both to standard linear model assumptions and the more flexible Generalized Beta regression. Of the 347 different ETGs, the nonparametric regression outperformed both the standard linear and generalized beta regression on all but 11. 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.

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 adjust the models appropriately. For example, the assumption of Gaussianity could be addressed by accounting 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 efficacy of Bayesian nonparametric regression in modeling 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 1990s (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).

Bayesian nonparametric regression can be used in all insurance applications where regression is used, and will be especially useful when the marginal distributions of the data given the predictors are non-Gaussian. We use episode treatment groups, which will be described in Section 2, to illustrate how healthcare data does not always meet the Gaussianity requirement for normal linear regression, and is in fact hard to classify in any distributional family (Huang et al., 2017). As will be shown, even the Generalized Beta distribution often falls short in characterizing healthcare costs, even though it is known for its flexibility (McDonald and Xu, 1995).

The specific ability of nonparametric regression to match any distributional shape will significantly improve model accuracy.

* Corresponding author.
E-mail addresses: richardson@stat.byu.edu (R. Richardson), hartman@stat.byu.edu (B. Hartman).

https://doi.org/10.1016/j.insmatheco.2018.06.002
0167-6687/

This will lead to more accurate predictions in forecasting that can be used for pricing. This will also be valuable in assessing the risk of future obligations. One area where the Bayesian nonparametric regression models significantly outperform all the others is in accurately capturing the tail behavior, which is commonly used in reserving and risk assessments.

Section 2 introduces the ETG dataset. Section 3 provides details for the dependent Dirichlet process ANOVA model used for Bayesian nonparametric regression. The ETG data analysis is shown in Section 4 with concluding remarks given in Section 5.

2. Data

Episode Treatment Groups (ETG) is a classification scheme for a variety of conditions that require medical services. ETGs are used 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.

Each record has two covariates, age and gender, 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 a certain demographic, such as pregnancy.

As stated in the introduction, an assumption of linear regression models is that the conditional distribution is either Gaussian, or whatever distributional assumption is used in the regression setting. Fig. 1 shows a marginal distribution for individuals in the dataset that are male and age 40. These marginal distributions under a Gaussian regression setting should be Gaussian, and they are clearly not. The form of non-Gaussianity is also inconsistent, meaning that simply choosing an alternative distributional assumption for the regression will also fall short. Some have heavy tails in one or both directions. Others are lightly or heavily skewed. The ETGs chosen to display are ones that are relatively specific in the diagnosis. The nature of the non-Gaussianity can become even more extreme in ETGs that are more broad, such as trauma to ear/nose/throat, dermatological signs and symptoms, and neurological diseases.

A number of insurance applications, such as pricing and risk assessment, will rely on accurate distributional assumptions based on data that looks very similar to these shown. Two different distributions with different tail behaviors may have similar mean and variances, but looking at value-at-risk or conditional tail expectation, where the tails are very important, will yield very different results. A modeling technique, such as Bayesian nonparametric regression, that is able to capture any conditional distribution behavior is a possible solution to accurate modeling of these healthcare costs.

Note that in pricing applications, the exact ETG will not be known in advance. The study we will conduct is conditional on knowing the exact ETG, so the methods here can be used for pricing, but only when combined with a separate study on expected ETGs or some distribution or knowledge of ETG frequency.

3. Bayesian nonparametric regression

A regression model with a single covariate could be written as

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

A flexible regression model could be constructed by isolating and flexibly modeling either the mean function, f, or the error distribution of ϵ_i. Fully nonparametric Bayesian regression, as opposed to assuming a specific form for the mean function or distributional family for the error, allows for distribution of $y_i|x_i$ to take any possible form in the specified domain. Functionally, this is done by using an infinite mixture of more standard relationships. The general idea is to apply a dependent Dirichlet process (DDP) (MacEachern, 1999) to the joint parameter space of the regression coefficients.

3.1. Dirichlet process

A standard building block of Bayesian nonparametric modeling is the Dirichlet process (DP). The formal definition defines a Dirichlet probability distribution for all partitions of a domain. A common functional equivalent is given by Sethuraman (1994), and is often referred as either the constructive definition of the DP or stick-breaking. This involves two components: a base distribution, G_0, from which draws $\theta_1, \theta_2, \ldots$ arise and weights $w_1, w_2, \ldots$ that follow the construction $w_l = \xi_l \prod_{i=1}^{l-1}(1-\xi_i)$ and $\xi_l \overset{i.i.d.}{\sim} Beta(1, \alpha)$. Then a Dirichlet Process, G, can be written as

$$G = \sum_{l=1}^{\infty} w_l \delta_{\theta_l}$$

where the function δ_{θ_l} returns a value of 1 at θ_l and 0 otherwise.

Especially using the constructive definition, it can be seen that the DP is a discrete distribution. To use DPs in a continuous setting, they are often used in conjunction with a mixture of normals, $f(x) = \int \phi(x|\mu, \sigma^2) dG(\mu)$, where $G(\mu)$ is a DP on the parameter space of μ and ϕ is a normal density function. Combining this with the constructive definition, a Dirichlet process mixture of normals can be written as

$$f(x) = \sum_{l=1}^{\infty} \phi(x|\mu_l, \sigma^2)$$

where μ_1, μ_2 come from the base distribution G_0 and $w_1, w_2, \ldots$ arise from stick-breaking. For this specification, σ^2 is the constant variance for each component of the mixture.

Intuitively, one can think of a DP mixture as an infinite mixture of normal distributions, but structured in such a way that it has complete support in the space of probability distributions. Conceptually, this means that whatever shape a distribution takes, if you use enough normal components in the mixture, you will be able to create an equivalent shape using a DP mixture. More information on this can be found in Gelman et al. (2013) and Müller et al. (2016) for those that wish to understand all the features of DPs and DP mixtures.

3.2. Dependent Dirichlet process

A dependent Dirichlet process (DDP) 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(α, $G_{0,D}$). 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 base 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 $\xi_l \overset{i.i.d.}{\sim} Beta(1, \alpha)$. 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, G_0.

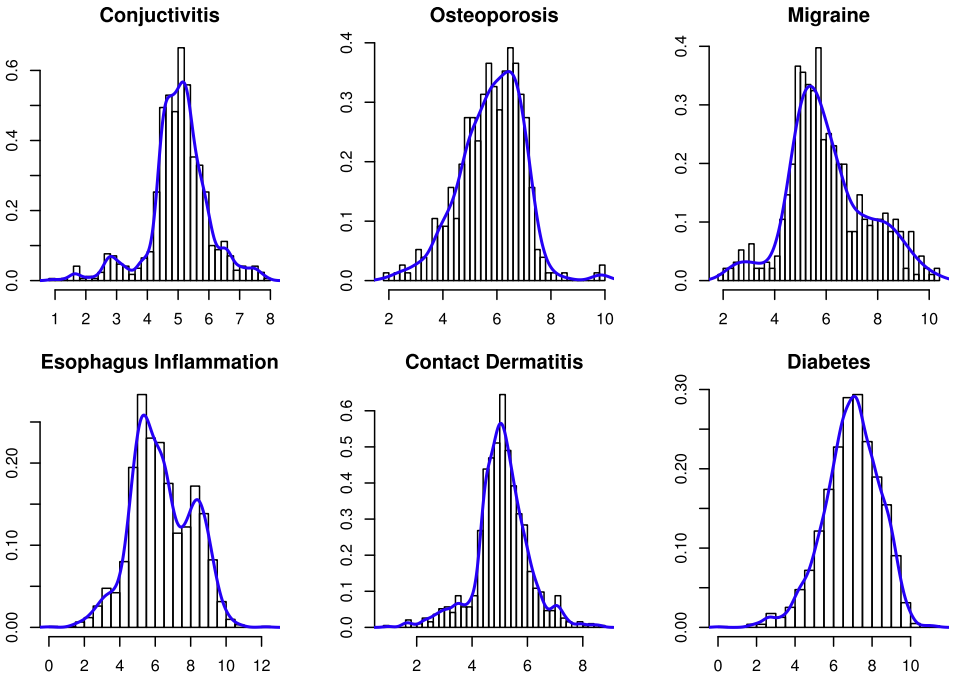

Fig. 1. The histogram of the log of the total charges for a male at age 40 for different episode treatment groups with a smoothed empirical density estimate overlaid.

The distribution of a finite number of points $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ 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 $(f(x_1), f(x_2), \ldots, f(x_n)) \sim G_{\mathbf{x}}$ 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 $\mathbf{y} = (f(x_1), f(x_2), \ldots, f(x_n))$ then

$$g(\mathbf{y}) = \int \phi(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) dG_{\mathbf{x}}(\boldsymbol{\mu}),$$

where $\boldsymbol{\mu} = (\mu(x_1), \mu(x_2), \ldots, \mu(x_n))$ 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

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

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 $G_{0,\mathbf{x}}$.

Just as a DP mixture can be thought of as a mixture of normal distributions, a DDP mixture can be functionally described as an infinite mixture of Gaussian processes, and for a finite number of points on the domain, the joint distribution is functionally an infinite mixture of multivariate normal distributions. This representation provides full support on the space of stochastic processes and multivariate densities, meaning that it can create an distributional characteristics with enough components in the mixture. More can be found in Müller et al. (2016) and MacEachern (2000).

3.3. 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, $\mathbf{z}_i = (1, z_{i,1}, z_{i,2}, \ldots, z_{i,p})'$. Then a DDP ANOVA model for $\mathbf{y} = (y_1, \ldots, y_n)'$ is

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

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 σ_l^2 and $\boldsymbol{\beta}_l$ are *a priori* independent. The base distribution will also have parameters ψ that can have hyperprior, $\pi(\psi)$.

The result of this construction of a regression model is flexible relationships between the covariates and the dependent 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.

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

where IG represents an inverse gamma distribution and NIW represents 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 η_b. The parameter α 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 results of the analysis, so they must be chosen carefully. μ_0 and Ψ_0 must be chosen to represent prior belief in the regression coefficients and the covariance, where κ_0 and ν_0 are chosen to represent the respective confidence in the prior belief of μ_0 and Ψ_0. The gamma hyperpriors could be chosen to yield expected values for a_σ and b_σ that represents belief in the variance of the regression model.

To aid in choosing priors that fit with interpretations, the dependent and independent variables 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. When the data is standardized, a 1 standard deviation increase in the ith independent variable is expected to change the dependent variable by β_i standard deviations, where β_i is the corresponding regression coefficient. Results will not be affected by this transformation as they can easily be transformed back to the original scale to make any specific inference. It has also been determined that this reduces the possibility of encountering computational errors.

3.4. Posterior inference

For a regression problem with n observations and p predictor variables, a fully nonparametric approach to Bayesian regression can be achieved through the model in Eqs. (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 Eq. (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 is found in Appendix B.

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.

3.5. Computational considerations

Bayesian nonparametric regression is more intensive computationally than standard linear regression. The basic structure of BNP regression is a weighted mixture of L regression equations, and the Gibbs' sampler algorithm that learns the parameters calculates a least squares-type formula iteratively for thousands of iterations, so it is easy to see that it can take a considerable amount of time. Additionally, as data size increases, BNP regression scales at the same rate as linear regression. For example, if there are n observations and p predictors, linear regression may take 1 s and BNP regression might take 1000 s. If you double the number of observations and linear regression takes 3 s, then BNP regression would take 3000. Similarly, if you double the number of predictors and linear regression takes 4 s, then BNP regression would take 4000.

BNP regression is also as flexible and straightforward as linear regression when adding more observations or predictors, transforming data points, adding polynomial structure, binning, or any other adjustment that can be made to linear regression.

4. Modeling claims by episode treatment group

To display the breadth of the advantage Bayesian nonparametric regression affords, we will analyze the data individually for all 347 ETGs. However, because the datasets can be quite large, only subsets are used in most cases. Individually, the computational burden of a single dataset is not too much for a standard machine to fit, but because we are doing all 347, using subsets considerably sped up the process. An analysis of a complete large dataset is shown in 4.2. Bayesian nonparametric regression, a standard Bayesian linear model, and an additional form on nonlinear regression using the Generalized Beta 2 (GB2) distribution 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 1000 data points were chosen at random, or all the data points were used if the size of the group was less than 1000. Age as a continuous variable and gender as a factor variable are used as the predictors because those are the only covariates in our data. In the cases where gender is incredibly skewed in favor of one gender, for example breast cancer, gender was left out as a covariate and only age was used. The DDP ANOVA model as well as a Bayesian linear regression model was fit to each dataset. 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, implying that 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 100.

The Bayesian linear 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 σ_2. The particular parameterization of the GB2 used, as shown in Frees and Valdez (2008), is

$$g(y_i) = \frac{e^{\alpha_1 \gamma_i}}{y_i |\sigma| B(\alpha_1, \alpha_2)[1 + e^{\gamma_i}]^{\alpha_1 + \alpha_2}} \tag{5}$$

where $\gamma_i = (\ln(y_i) - \mu_i)/\sigma$. For regression, we set $\mu_i = \boldsymbol{z}_i' \boldsymbol{\beta}$. Then the parameters to estimate in this model for the ETG datasets are $\beta_0, \beta_1, \beta_2, \sigma, \alpha_1$, and α_2. For the GB2 regression, the parameters have significantly different interpretations, so the same priors could not be used as the BNP or Gaussian models. Specifically, σ is given a Gamma prior with $\alpha = 2$ and $\beta = 0.1$, α_1 and α_2 are given independent Gamma priors with $\alpha = 4$ and $\beta = 4$, and the regression coefficients have normal priors with a mean of 0 and a standard deviation of 1000. The priors are intentionally diffuse, except for the α parameters where diffuse priors made them difficult to estimate.

The metric used to assess model fit is Deviance Information Criterion (DIC). It is analogous to Akaike'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. DIC accounts for distributional accuracy and penalizes for higher complexity, which means that a model with more parameters would get penalized more than a model with fewer parameters. This metric seems the most appropriate for comparing the BNP regression to the GB2 regression and normal linear model as it will be able to detect if the more flexible fit of the BNP regression is worth the additional complexity.

In all 347 ETGs, both the BNP and GB2 regression models outperformed the Gaussian linear model. Additionally, the BNP model

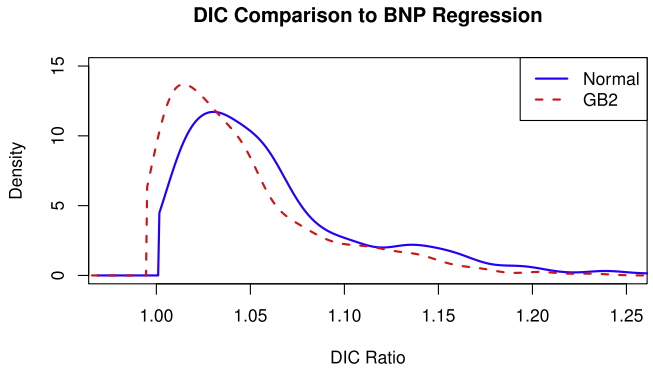

Fig. 2. The ratio of DIC between the Gaussian to the BNP regression and the GB2 to the BNP regression models are shown as a density for the 347 datasets.

outperformed the GB2 model in all but 11 of the ETGs, suggesting that the additional complexity in the BNP model is valuable when predicting healthcare costs. Fig. 2 shows the DIC ratios between the Gaussian and GB2 regression models to the BNP regression model. Any ratio above 1 implies that the BNP outperformed the other model. It shows that the DIC is consistently lower for the BNP. A common characteristic of datasets where the DIC was similar between the three models was smaller datasets with few outliers.

We can assess the quality of the posterior predictive samples by using a continuous rank probability score (CRPS) to evaluate prediction accuracy (Gneiting and Raftery, 2007). If $y^{(1)}, \ldots, y^{(B)}$ are B samples from a fitted distribution and y_T is an observation, then CRPS is

$$CRPS_y = \frac{1}{n}\sum_{i=1}^{B}(y^{(i)} - y_T)^2 + \frac{1}{n^2}\sum_{i=1}^{B}\sum_{j=1}^{B}(y^{(i)} - y^{(j)})^2 \tag{6}$$

CRPS is better at assessing distributional accuracy of predictions then mean square prediction error, which only accounts for the point estimate.

To calculate CRPS, we use 100 values that were left out of the subset, or 10% of the data if the number of observations was less than 1000, to find out of sample prediction accuracy. The average CRPS for linear regression was 2.146. The average CRPS for BNP regression was 1.868, suggesting that the BNP regression did significantly better in out of sample predictions.

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 Fig. 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, GB2 regression, and DDP ANOVA model were fit to the training dataset, which comprised 90% of the data, a total of 160,228 observations. The other 10% was left out to predict. The median values and 95% credible intervals for the coefficients using the Bayesian linear model are shown in Table 1 along with the average effects from the nonparametric regression model, where $\hat{\beta}_1$ 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 1 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 1
Median and 95% credible interval for the sampled coefficients for the BLM and the average effects of the BNP regression model.

		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
DIC values for the three models for the Conjunctivitis ETG.

Model	DIC
BLM	12,600
GB2	12,300
BNP	10,600

Table 3
Some percentiles of the fitted distribution for an 18 year old female for all three models.

Percentile	25	50	75
BLM	73.0	135.3	246.9
GB2	84.7	135.3	207.8
BNP	96.8	140.6	217.4
Test data	93.4	144	217.6

Table 2 lists the DIC values for the three model fits. The posterior predictive distribution is plotted for both models for an 18 year old female in Fig. 4. The posterior predictive distributions are overlaid the histogram of all 18 year females in the test dataset. The extra bump in the left tail is accurately captured by the posterior predictive distribution of the BNP regression model. To try and capture the tail, the BLM model gives a wider prediction interval than is necessary.

Table 3 shows some percentiles of the predictive distribution as well, showing once again the accuracy of the BNP model. Predictive distributions for other ages and for males showed similar trends. Both the DIC results and the predictive distributions show that the BNP regression model is able to capture a feature of the data that gives it a significant advantage over the other models.

4.3. Lung transplant

Lung transplant was also chosen to examine a smaller dataset more carefully. 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 GB2 and BNP models will predict thicker tails in the posterior predictive distribution. The BLM model will just give 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 Fig. 5. The wide variance in the BLM predictions can be seen. The thicker tails in the non-Gaussian regression models are difficult to detect by eye, but they are thicker than the tails of a standard Gaussian. In all 4 cases, the actual observation, which, again, was not used in the model fit, is within the bounds of both prediction intervals. This actual observation is shown in the plots by a vertical spike.

As mentioned in the section with the results of all 347 ETGs, datasets with a smaller number of data points typically saw less of an advantage from the more flexible regression models. Considering the values for DIC given in Table 4, the GB2 does better for this ETG, but not by much compared to the BNP regression. The Gaussian linear model performs the worst, most likely due to the outliers.

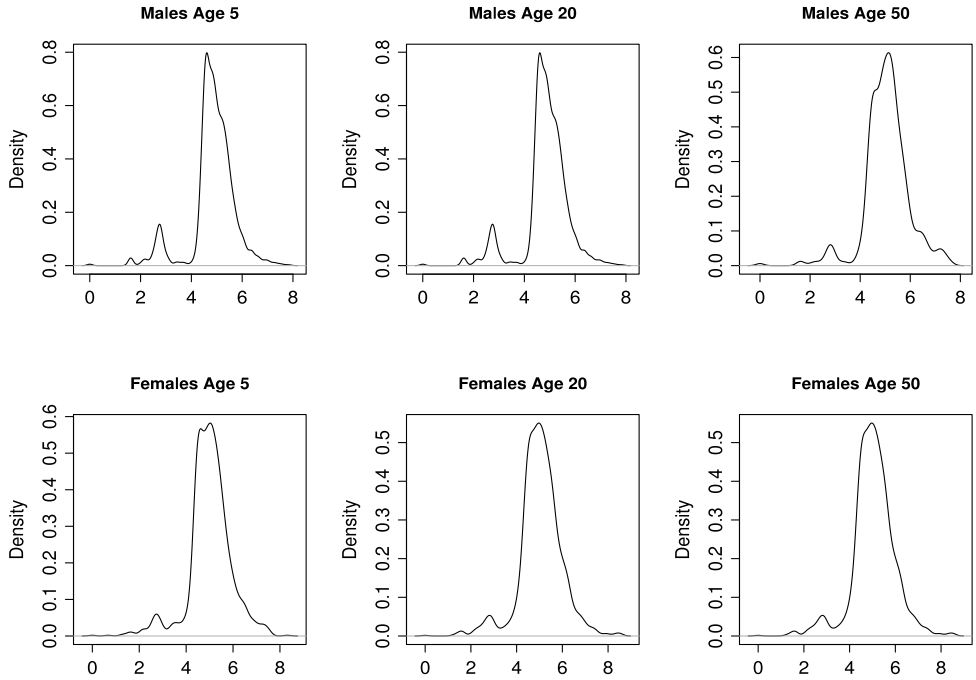

Fig. 3. Empirical densities of the total charges for the conjunctivitis ETG for a number of covariate combinations.

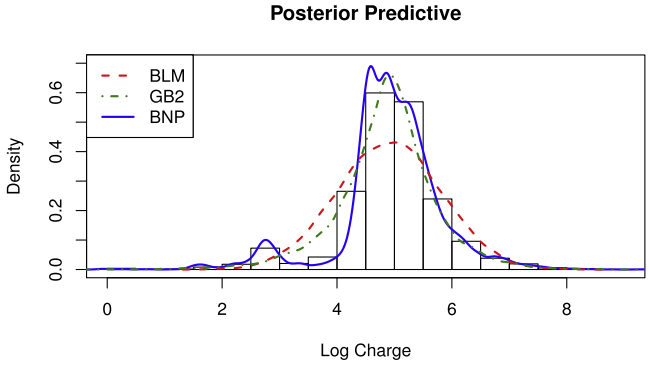

Fig. 4. The posterior predictive distribution for the log charges of an 18 year old female with conjunctivitis overlaid the histogram of the data of all 18 year females in the ETG dataset.

5. Conclusion

Bayesian non-parametric regression can be applied generally everywhere regression is used in insurance applications. As stated previously, the specific case study used here assumes knowledge of the ETGs. For this example to be used directly in pricing, it would need to also include either a separate model or additional information of ETG frequency. This study could also be especially useful for reserving, considering the improved distributional fit.

The utility of a non-Gaussian regression relationship has been illustrated for healthcare data in a number of previous papers. This paper has shown that Bayesian non-parametric regression is a very powerful tool for flexible regression modeling of healthcare claim data and vastly performs both linear and GB2 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).

Table 4
DIC values for the three models for the Lung Transplant ETG.

Model	DIC
BLM	1080
GB2	1002
BNP	1005

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 Section 3, 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.

Appendix 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 Eq. (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 is needed for the data. For a specific value of α the approximate expected value of N is $E(N|\alpha) \approx \alpha \log\left(\frac{\alpha+n}{\alpha}\right)$ and the approximate

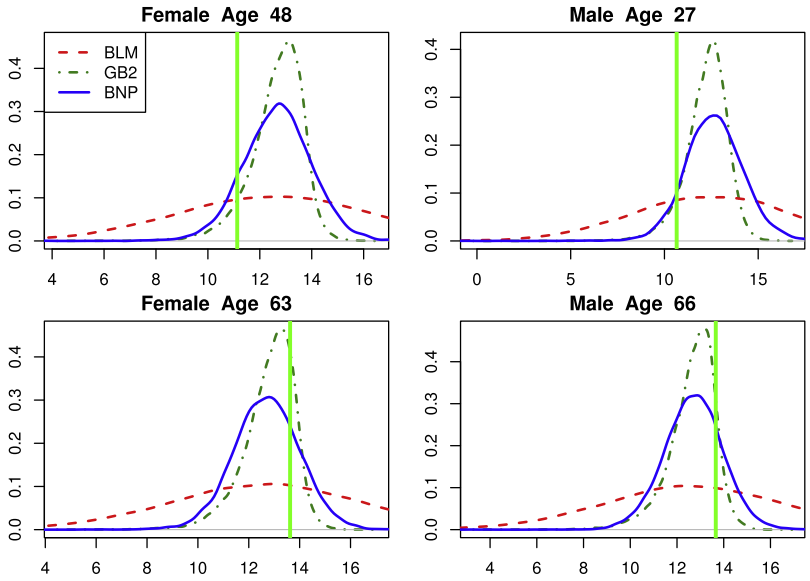

Fig. 5. The posterior predictive distribution for the log charges of 4 individuals left out of the analysis for prediction for the Lung Transplant ETG. The solid vertical lines are the observed values.

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 α 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, $L_1, \ldots, L_n$. These can take in integer values between 1 and N, assigning each observation to one of the N clusters. Eqs. (1) and (2) in this case can be rewritten as

$$y_i|\mathbf{z}_i, L_i \sim N(y_i; \mathbf{z}_i'\boldsymbol{\beta}_{L_i}, \sigma^2_{L_i}) \tag{A.1}$$

$$L_i|\mathbf{w} \sim \sum_{i=1}^{N} w_l \delta_l(L_i). \tag{A.2}$$

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 $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N$, (2) the N variance atoms, $\sigma_1^2, \ldots, \sigma_N^2$, (3) the weights $w_1, \ldots, w_N$, (4) the assignment variables $L_1, \ldots, L_n$, (5) the hyperpriors μ_β, Σ_β, a_σ, and b_σ, and (6) the value for α.

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

$$\boldsymbol{\beta}_l|\cdot \sim N(\boldsymbol{\beta}_l|\mu^*_\beta, \Sigma^*_\beta) \tag{A.3}$$

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

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

$$\sigma_l^2|\cdot \sim \text{IG}\left(\sigma_l^2|a_\sigma + M_l/2, b_\sigma + .5\left(\sum_{L_i=l} y_i^2 + \mu'_\beta\Sigma_\beta^{-1}\mu_\beta + \mu'^*_\beta\Sigma_\beta^{*-1}\mu^*_\beta\right)\right) \tag{A.4}$$

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 $l = 1, \ldots, N-1$. Then $w_l = \xi_l \prod_{i=1}^{l-1}(1-\xi_i)$ for $l = 1, \ldots, N-1$ 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 μ_β and Σ_β, posterior samples can be taken from

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

The posterior samples for b_σ 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_σ. It can be drawn using a Metropolis–Hastings algorithm, where a proposal is made for a new value of a_σ given the previous value. If this generating distribution is $g(a_\sigma^*|a_\sigma)$ then the new value a_σ^* 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_α, b_α) prior, posterior samples for α 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.

Appendix 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 $b = 1, \ldots, B$, meaning variables superscripted by (b) are the bth 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 σ^{*2} equal to $\sigma_{l^*}^2$
3. Draw a new value $y^{*(b)}$ from $N(\cdot|\mathbf{z}^{*'}\boldsymbol{\beta}^*, \sigma^{*2})$

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

References

De Iorio, M., Johnson, W.O., Müller, P., Rosner, G.L., 2009. Bayesian nonparametric nonproportional hazards survival modeling. Biometrics 65 (3), 762–771.

De Iorio, M., Müller, P., Rosner, G.L., MacEachern, S.N., 2004. An ANOVA model for dependent random measures. J. Amer. Statist. Assoc. 99 (465), 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 Anal. 10 (4), 821–847.

Dunson, D.B., Xing, C., 2009. Nonparametric Bayes modeling of multivariate categorical data. J. Amer. Statist. Assoc. 104 (487), 1042–1051.

Fellingham, G.W., Kottas, A., Hartman, B.M., 2015. Bayesian nonparametric predictive modeling of group health claims. Insurance Math. Econom. 60, 1–10.

Frees, E.W., Valdez, E.A., 2008. Hierarchical insurance claims modeling. J. Amer. Statist. Assoc. 103 (484), 1457–1469.

Gan, G., Valdez, E.A., 2017. Fat-tailed regression modeling with spliced distributions, Available at SSRN: https://ssrn.com/abstract=3037062.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian Data Analysis. CRC press.

Gneiting, T., Raftery, A.E., 2007. Strictly proper scoring rules, prediction, and estimation. J. Amer. Statist. Assoc. 102 (477), 359–378.

Hong, L., Martin, R., 2017. A flexible Bayesian nonparametric model for predicting future insurance claims. N. Am. Actuar. J. 21 (2), 228–241.

Huang, S., Hartman, B., Brazauskas, V., 2017. Model selection and averaging of health costs in episode treatment groups. ASTIN Bull.: J. IAA 47 (1), 153–167.

Jara, A., Hanson, T.E., Quintana, F.A., Müller, P., Rosner, G.L., 2011. DPpackage: Bayesian semi-and nonparametric modeling in R. J. Stat. Softw. 40 (5), 1.

Kottas, A., Müller, P., Quintana, F., 2005. Nonparametric Bayesian modeling for multivariate ordinal data. J. Comput. Graph. Statist. 14 (3), 610–625.

Kudryavtsev, A.A., 2009. Using quantile regression for rate-making. Insurance Math. Econom. 45 (2), 296–304.

MacEachern, S.N., 1999. Dependent nonparametric processes. In: ASA Proceedings of the Section on Bayesian Statistical Science, Vol. 1999. Alexandria, Virginia. Virginia: American Statistical Association, pp. 50–55.

MacEachern, S.N., 2000. Dependent dirichlet processes. In: Unpublished Manuscript. Department of Statistics, the Ohio State University, pp. 1–40.

McDonald, J.B., Xu, Y.J., 1995. A generalization of the beta distribution with applications. J. Econometrics 66 (1–2), 133–152.

Miljkovic, T., Grün, B., 2016. Modeling loss data using mixtures of distributions. Insurance Math. Econom. 70, 387–396.

Müller, P., Erkanli, A., West, M., 1996. Bayesian curve fitting using multivariate normal mixtures. Biometrika 83 (1), 67–79.

Müller, P., Quintana, F.A., Jara, A., Hanson, T., 2016. Bayesian Nonparametric Data Analysis. Springer.

Neter, J., Kutner, M.H., Nachtsheim, C.J., Wasserman, W., 1996. Applied Linear Statistical Models, Vol. 4. Irwin Chicago.

Rousseeuw, P., Daniels, B., Leroy, A., 1984. Applying robust regression to insurance. Insurance Math. Econom. 3 (1), 67–72.

Sethuraman, J., 1994. A constructive definition of Dirichlet priors. Statist. Sinica 639–650.

Shi, P., 2013. Fat-tailed regression models. In: Predictive Modeling Applications in Actuarial Science, Vol. 1. pp. 236–259.