Bayesian Nonparametric Predictive Modeling of Group Health Claims

Gilbert W. Fellingham[a,*], Athanasios Kottas[b], Brian M. Hartman[c]

[a]Brigham Young University
[b]University of California, Santa Cruz
[c]University of Connecticut

Abstract

Models commonly employed to fit current claims data and predict future claims are often parametric and relatively inflexible. An incorrect model assumption can cause model misspecification which leads to reduced profits at best and dangerous, unanticipated risk exposure at worst. Even mixture models may not be sufficiently flexible to properly fit the data. Using a Bayesian nonparametric model instead can dramatically improve claim predictions and consequently risk management decisions in group health practices. The improvement is significant in both simulated and real data from a major health insurer's medium-sized groups. The nonparametric method outperforms the hierarchical method, especially when predicting future claims for new business. In our sample, the nonparametric model outperforms the hierarchical model in 84% of the renewal business and 88% of the new business. This is particularly important in light of the healthcare reform in the United States and as healthcare costs rise around the world.

Keywords: Dirichlet process prior, multimodal prior, prediction

1. Introduction

As George Box famously said, "essentially, all models are wrong, but some are useful" (Box and Draper, 1987). This is especially true when the process being modeled is either not well understood or the necessary data are

[*]Corresponding Author: 223H TMCB, Provo, UT 84602, USA
Phone:(801)422-2806 Fax:(801)422-0635 email:gwf@byu.edu

unavailable. Both are concerns in health insurance. Our knowledge of the human body and understanding of what makes it sick are limited, but the main difficulty is lack of available data; limited by both technology/cost (e.g. DNA sequences and complete blood panels) and privacy (e.g. patient records especially of prospective policyholders). This is even more prevalent in group health where data on the individual policyholders can be sparse. Bayesian nonparametric (BNP) models are a flexible option to describe both current and prospective healthcare claims. As will be shown, in modeling group health claims BNP models are superior to traditional Bayesian parametric models. Both model types could be used in premium calculations for small groups or prospective blocks of business, and to calculate experience-based refunds. Precise estimation is especially important now with the introduction of the Affordable Care Act in the United States and as healthcare costs continue to consume an increasing share of personal wealth around the world

One of the principles of Bayesian methods very familiar to actuaries is improvement in the process of estimating, say, the pure premium for a block of business by "borrowing strength" from related experience through credibility. For example, if the size of a block is small enough, the exposure in previous years may be limited. In this case, estimates of future costs may be based more heavily on other, related experience in an effort to mitigate the effects of small sample random variation.

Hierarchical Bayesian models offer an extremely useful paradigm for prediction in this setting. However, in somewhat simplistic terms, successful Bayesian model specification hinges on selecting scientifically appropriate prior distributions. When there is an unanticipated structure in the function defining the prior, posterior distributions (and prediction) will, by definition, be flawed.

This leads us to consider the Bayesian nonparametric formulation. Bayesian nonparametric modeling specifies distributions on function spaces. An increased probability of obtaining more precise prediction comes with the increased flexibility of BNP methods. In this paper, we will demonstrate why BNP models are useful when building statistical models, especially when prediction is the primary inferential objective.

A brief outline of the paper follows. First, we specify the mathematical structure of the models in the full parametric and nonparametric settings. The parametric model is described first since the nonparametric setting parallels and extends the parametric setting. We provide more detail for the nonparametric setting since it is less familiar. Additionally, we provide the algorithms necessary to implement the nonparametric model in Appendix B. We next present a small simulation study to demonstrate the performance of the two models in situations where the structure used to generate the data is known. Finally, we present results from analyses of claims data from 1994 and compare the two formulations by evaluating their performance in predicting costs in 1995.

2. The models

2.1. The hierarchical parametric Bayes model

We present the traditional parametric Bayesian model first since the nonparametric formulation is based on the parametric version. To develop the parametric model, we need to characterize the likelihood and the prior distributions of the parameters associated with the likelihood. There are two things to consider when thinking about the form of the likelihood: the probability a claim will be made and the amount of the claim, given a claim is made. The probability a claim is made differs from group to group and in our data is around 0.70. Thus, about 30% of the data are zeros, meaning no claim was filed for those particular policies. We chose to deal with this by using a likelihood with a point mass at zero with probability π_i for group i. The parameter π_i depends on the group membership.

The cost of a claim given that a claim is paid is positively skewed. We choose a gamma density for this portion of the likelihood with parameters γ and θ. In a previous analysis of this data, Fellingham et al. (2005, p. 11) indicated that "the gamma likelihood for the severity data is not rich enough to capture the extreme variability present in this type of data." However, we will show that with the added richness furnished by the nonparametric model, the gamma likelihood is sufficiently flexible to model the data.

Let $f(y;\gamma,\theta)$ denote the density at y of the gamma distribution with shape parameter γ and scale parameter θ. Hence,

$$f(y;\gamma,\theta) = \frac{1}{\theta^\gamma \Gamma(\gamma)} y^{\gamma-1} \exp\left(\frac{-y}{\theta}\right). \tag{1}$$

The likelihood follows using a compound distribution argument:

$$\prod_{i=1}^{N_g} \prod_{\ell=1}^{L_i} \Big[\pi_i I(y_{i\ell}=0) + (1-\pi_i) f(y_{i\ell};\gamma_i,\theta_i) I(y_{i\ell}>0)\Big], \tag{2}$$

where i indexes the group number, N_g is the number of groups, ℓ indexes the observation within a specific group, L_i is the number of observations within group i, π_i is the proportion of zero claims for group i, θ_i and γ_i are the parameters for group i, $y_{i\ell}$ is the cost per day of exposure for each policyholder, and I denotes the indicator function. Thus, we have a point mass probability for $y_{i\ell} = 0$ and a gamma likelihood for $y_{i\ell} > 0$.

As discussed in the opening section, the choice of prior distributions is critical. One of the strengths of the full Bayesian approach is the ability it gives the analyst to incorporate information from other sources. Because we had some previous experience with the data that might have unduly influenced our choices of prior distributions, we chose to use priors that were only moderately informative. These priors were based on information available for other policy types. We did not use any of the current data to make decisions about prior distributions. Also, we performed a number of sensitivity analyses in both the parametric and the nonparametric settings and found that the results were not sensitive to prior or hyperprior specification in either case.

For the first stage of our hierarchical prior specification, we need to choose random-effects distributions for the parameters π_i and (γ_i, θ_i). We assume independent components conditionally on hyperparameters. In particular,

$$\begin{aligned} \pi_i \mid \mu_\pi &\overset{\text{ind.}}{\sim} \text{Beta}(\mu_\pi, \sigma_\pi^2), \;\; i = 1, ..., N_g, \\ \gamma_i \mid \beta &\overset{\text{ind.}}{\sim} \text{Gamma}(b, \beta), \;\; i = 1, ..., N_g, \\ \theta_i \mid \delta &\overset{\text{ind.}}{\sim} \text{Gamma}(d, \delta), \;\; i = 1, ..., N_g. \end{aligned} \tag{3}$$

Here, to facilitate prior specification, we work with the Beta distribution parametrized in terms of its mean μ_π and variance σ_π^2, that is, with density given by

$$\frac{1}{\text{Be}(c_1, c_2)} \pi^{c_1 - 1}(1 - \pi)^{c_2 - 1}, \;\; \pi \in (0, 1), \tag{4}$$

where $c_1 = \sigma_\pi^{-2}(\mu_\pi^2 - \mu_\pi^3 - \mu_\pi \sigma_\pi^2)$, $c_2 = \sigma_\pi^{-2}(\mu_\pi - 2\mu_\pi^2 + 3\mu_\pi^3 - \sigma_\pi^2 + \mu_\pi \sigma_\pi^2)$, and $\text{Be}(\cdot, \cdot)$ denotes the Beta function, $\text{Be}(r, t) = \int_0^1 u^{r-1}(1-u)^{t-1} \text{d}u$, $r > 0$, $t > 0$ (Forbes et al., 2011). We choose specific values for the hyperparameters σ_π^2, b, and d and assign reasonably non-informative priors to μ_π, β and δ. We note that sensitivity analyses showed that the values chosen for the hyperparameters had virtually no impact on the outcome. For the prior distributions, we take a uniform prior on $(0, 1)$ for μ_π and inverse gamma priors for β and δ with shape parameter equal to 2 (implying infinite prior variance) and scale parameters A_β and A_δ, respectively. Hence, the prior density for β is given by $A_\beta^2 \beta^{-3} \exp(-A_\beta/\beta)$ (with an analogous expression for the prior of δ). Further details on the choice of the values for σ_π^2, b, d, A_β, and A_δ in the analysis of the simulated and real data are provided in Sections 4 and 5, respectively.

The posterior for the full parameter vector

$$(\{(\pi_i, \gamma_i, \theta_i) : i = 1, ..., N_g\}, \mu_\pi, \beta, \delta)$$

is then proportional to

$$\left[\prod_{i=1}^{N_g} \frac{\beta^{-b}}{\Gamma(b)} \gamma_i^{b-1} \exp(\frac{-\gamma_i}{\beta}) \frac{\delta^{-d}}{\Gamma(d)} \theta_i^{d-1} \exp(\frac{-\theta_i}{\delta}) \frac{1}{\mathrm{Be}(c_1, c_2)} \pi_i^{c_1 - 1} (1-\pi_i)^{c_2 - 1}\right] \\ \times \left[\prod_{i=1}^{N_g} \prod_{\ell=1}^{L_i} \{\pi_i I(y_{i\ell} = 0) + (1-\pi_i) f(y_{i\ell}; \gamma_i, \theta_i) I(y_{i\ell} > 0)\}\right] p(\mu_\pi) p(\beta) p(\delta), \tag{5}$$

where $p(\mu_\pi)$, $p(\beta)$, and $p(\delta)$ denote the hyperpriors discussed above.

This model can be analyzed using Markov chain Monte Carlo (MCMC) to produce samples from the posterior distributions of the parameters (Gilks et al., 1995). To predict new data, we first draw new parameter values by using the marginalized version of the model obtained by integrating over the hyperprior distributions. Operationally, this means taking the current values of the hyperparameters at each iteration of the MCMC and drawing values of the $(\gamma_*, \theta_*, \pi_*)$ from their respective prior distributions given the current values of the hyperparameters. Predicted values are then drawn from the likelihood using $(\gamma_*, \theta_*, \pi_*)$. Prediction of new data is therefore dependent on the form of the prior distributions of the parameters. The importance of this idea cannot be overstated. The consequence of this notion is that if the prior distributions are misspecified, draws of new parameters will not mirror the actual setting, and predictions of new data must be incorrect. Estimation of parameters present in the current model will not be impacted as long as the prior distributions have appropriate support and are not so steep as to overpower the data. The impact on estimating costs is that those costs arising from groups that may be present in the future but are not being modeled with the current data must be wrong if the prior specification of the parameters' distribution is not accurate. This reveals the strength of the nonparametric model. Since the nonparametric prior is placed on the space of *all* plausible functions rather than on the parameters of a single function, the appropriate prior specification will be uncovered during the analysis. We demonstrate the impact of this idea in Section 5.

2.2. The nonparametric Bayesian model

The parametric random-effects distributions chosen for the π_i, γ_i, and θ_i in Section 2.1 might not be appropriate for specific data sets. Moreover, since these are distributions for latent model parameters, it is not intuitive to anticipate their form and shape based on exploratory data analysis. Bayesian nonparametric methods provide a flexible solution to this problem. The key idea is to use a nonparametric prior on the random-effects distributions that supports essentially all possible distribution shapes. That is, the nonparametric model allows the shape of the random-effects distributions to be driven by the data and to take any form. Since the nonparametric prior model can be centered around familiar parametric forms, it is still relatively simple to develop approaches to prior elicitation.

Thus, through the prior to posterior updating of BNP models, the data are allowed to drive the shape of the posterior random-effects distributions. This shape can be quite different from standard parametric forms (when these forms are not supported by the data), resulting in more accurate posterior predictive inference when using the nonparametric formulation.

Here, we utilize Dirichlet process (DP) priors, a well-studied class of nonparametric prior models for distributions which allow the data to drive the shape of the posterior predictive distributions. We refer the interested reader to Appendix A for a brief review of Dirichlet processes. For a more extensive review, see also Dey et al. (1998); Hanson et al. (2005); Müller and Quintana (2004); Walker et al. (1999).

We formulate a nonparametric extension of the parametric model discussed in the previous section by replacing the hierarchical parametric priors for the random-effects distributions with hierarchical DP priors (formally, mixtures of DP priors). The DP can be defined in terms of two parameters: a positive scalar parameter α, which can be interpreted as a precision parameter, and a specified base (centering) parametric distribution G_0.

While it would have been possible to place the DP prior on the joint random-effects distribution associated with the triple $(\gamma_i, \theta_i, \pi_i)$, that course of action would require that the parameters be updated as a group. Since it is possible that the probability of no claim being made is not associated with the distribution of costs within a group, we have chosen to treat these parameters separately. Thus, we have a DP prior for the random-effects distribution, G_1, which is associated with the π_i, as well as a separate (independent) DP prior for the random-effects distribution, G_2, which corresponds to the (γ_i, θ_i).

Now we have the following hierarchical version for the nonparametric model:

$$
\begin{aligned}
y_{i\ell} \mid \pi_i, \gamma_i, \theta_i &\overset{\text{ind.}}{\sim} \pi_i I(y_{i\ell} = 0) + (1 - \pi_i) f(y_{i\ell}; \gamma_i, \theta_i) I(y_{i\ell} > 0), \\
& \qquad \ell = 1, ..., L_i; \;\; i = 1, ..., N_g \\
\pi_i \mid G_1 &\overset{\text{i.i.d.}}{\sim} G_1, \;\; i = 1, ..., N_g \\
(\gamma_i, \theta_i) \mid G_2 &\overset{\text{i.i.d.}}{\sim} G_2, \;\; i = 1, ..., N_g \\
G_1, G_2 &\overset{\text{ind.}}{\sim} \text{DP}(\alpha_1, G_{10}) \times \text{DP}(\alpha_2, G_{20}).
\end{aligned}
\tag{6}
$$

Here, $\alpha_1, \alpha_2 > 0$ are the precision parameters of the DP priors, and G_{10} and G_{20} are the centering distributions. Again, the DP priors allow the distributions G_1 and G_2 to take flexible prior shapes. The precision parameters α_1 and α_2 control how close a prior realization G_k is to G_{k0} for $k = 1, 2$. But in the resulting posterior estimates, the distributional shape for G_1 and G_2 can assume nonstandard forms that may be suggested by the data, since we are not insisting that the prior model for G_1 and G_2 take on specific parametric forms such as the Beta and Gamma forms in equation (3). The importance of allowing this level of flexibility is illustrated with the analysis of the claims data in Section 5.

We set $G_{10}(\pi) = \text{Beta}(\pi; \mu_\pi, \sigma_\pi^2)$, which is the random-effects distribution used for the π_i in the parametric version of the model. Again, we place a uniform prior on μ_π and take σ_π^2 to be fixed. For G_{20} we take independent Gamma components, $G_{20}((\gamma, \theta); \beta, \delta) = \text{Gamma}(\gamma; b, \beta) \times \text{Gamma}(\theta; d, \delta)$, with fixed shape parameters b and d, and inverse gamma priors assigned to β and δ. Again, note that G_{20} is the random-effects distribution for the (γ_i, θ_i) used in the earlier parametric version of the model. In all analyses, we kept α_1 and α_2 fixed.

In the DP mixture model in (6), the precision parameters control the distribution of the number of distinct elements N_1^* of the vector $\{\pi_1, \ldots, \pi_{N_g}\}$ (controlled by α_1) and N_2^* of the vector $\{(\gamma_1, \theta_1), \ldots, (\gamma_{N_g}, \theta_{N_g})\}$ (controlled by α_2), and hence, the number of distinct components of the mixtures. The number of distinct groups is smaller than N_g with positive probability, and for typical choices of α_1 and α_2 is fairly small relative to N_g. For instance, for moderate to large N_g,

$$E(N_k^* \mid \alpha_k) \approx \alpha_k \log\left(\frac{\alpha_k + N_g}{\alpha_k}\right), k = 1, 2, \tag{7}$$

and exact expressions for the prior probabilities $\Pr(N_k^* = m \mid \alpha_k)$, $m = 1, \ldots, N_g$ are also available (e.g. Escobar and West, 1995). These results are useful in choosing the values of α_1 and α_2 for the analysis of any particular data set using model (6).

2.2.1. Posterior inference

To obtain posterior inference, we work with the marginalized version of model (6), which results from integrating G_1 and G_2 over their independent DP priors,

$$\begin{array}{rcl} y_{i\ell} \mid \pi_i, \gamma_i, \theta_i & \overset{\text{ind.}}{\sim} & \pi_i I(y_{i\ell} = 0) \\ & & +(1-\pi_i) f(y_{i\ell}; \gamma_i, \theta_i) I(y_{i\ell} > 0), \\ & & \ell = 1, ..., L_i; \;\; i = 1, ..., N_g \\ (\pi_1, ..., \pi_{N_g}) \mid \mu_\pi & \sim & p(\pi_1, ..., \pi_{N_g} \mid \mu_\pi) \\ (\gamma_1, \theta_1), ..., (\gamma_{N_g}, \theta_{N_g}) \mid \beta, \delta & \sim & p((\gamma_1, \theta_1), ..., (\gamma_{N_g}, \theta_{N_g}) \mid \beta, \delta), \\ \beta, \delta, \mu_\pi & \sim & p(\beta), p(\delta), p(\mu_\pi), \end{array} \tag{8}$$

where, as before, $p(\beta)$, $p(\delta)$, and $p(\mu_\pi)$ denote the hyperpriors for β, δ, and μ_π.

Key to the development of the posterior simulation method is the form of the prior for the π_i and for the (γ_i, θ_i) induced by the DP priors for G_1 and G_2 respectively. The joint prior for the π_i and for the (γ_i, θ_i) can be developed using the Pólya urn characterization of the DP (Blackwell and MacQueen, 1973). Specifically,

$$p(\pi_1, ..., \pi_{N_g} \mid \mu_\pi) = $$
$$g_{10}(\pi_1; \mu_\pi, \sigma_\pi^2) \prod_{i=2}^{N_g} \left\{ \frac{\alpha_1}{\alpha_1 + i - 1} g_{10}(\pi_i; \mu_\pi, \sigma_\pi^2) + \frac{1}{\alpha_1 + i - 1} \sum_{j=1}^{i-1} \delta_{\pi_j}(\pi_i) \right\}, \tag{9}$$

and $p((\gamma_1, \theta_1), ..., (\gamma_{N_g}, \theta_{N_g}) \mid \beta, \delta)$ is given by

$$g_{20}((\gamma_1, \theta_1); \beta, \delta) \times \prod_{i=2}^{N_g} \left\{ \frac{\alpha_2}{\alpha_2 + i - 1} g_{20}((\gamma_i, \theta_i); \beta, \delta) + \frac{1}{\alpha_2 + i - 1} \sum_{j=1}^{i-1} \delta_{(\gamma_j, \theta_j)}(\gamma_i, \theta_i) \right\}, \quad (10)$$

where g_{10} and g_{20} denote respectively the densities corresponding to G_{10} and G_{20}, and $\delta_a(y)$ denotes a point mass for y at a (i.e., $\Pr(y = a) = 1$ under the $\delta_a(\cdot)$ distribution for y). These expressions are key for MCMC posterior simulation, since they yield convenient forms for the prior full conditionals for each π_i and for each (γ_i, θ_i). In particular, for each $i = 1, ..., N_g$,

$$p(\pi_i \mid \{\pi_j : j \neq i\}, \mu_\pi) = \frac{\alpha_1}{\alpha_1 + N_g - 1} g_{10}(\pi_i; \mu_\pi, \sigma_\pi^2) + \frac{1}{\alpha_1 + N_g - 1} \sum_{j=1}^{N_g - 1} \delta_{\pi_j}(\pi_i) \quad (11)$$

and

$$p((\gamma_i, \theta_i) \mid \{(\gamma_j, \theta_j) : j \neq i\}, \beta, \delta) = \frac{\alpha_2}{\alpha_2 + N_g - 1} g_{20}((\gamma_i, \theta_i); \beta, \delta) + \frac{1}{\alpha_2 + N_g - 1} \sum_{j=1}^{N_g - 1} \delta_{(\gamma_j, \theta_j)}(\gamma_i, \theta_i). \quad (12)$$

Intuitively, the idea for posterior sampling using expressions (11) and (12) is that proposal values for the parameters are drawn from either the centering distribution (with probability $\alpha_k(\alpha_k + N_g - 1)^{-1}, k = 1, 2$) or from values for previous draws of the other parameters (with probabilities $(\alpha_k + N_g - 1)^{-1}$, for $j \neq i$, and with $k = 1, 2$). These proposal values are then treated as in the parametric setting and are either kept or rejected in favor of the current value for the parameter.

Implementation of the MCMC method to produce samples from the posterior distributions is not much more difficult than in the parametric setting. For specific details concerning implementation of the MCMC algorithm in this nonparametric model, we refer the interested reader to Appendix B.

2.2.2. Posterior predictive inference

We will focus on the posterior predictive distribution for a new group — that is, a group for which we have no data. The cost for a (new) policyholder within a new group is denoted by y_*. To obtain $p(y_* \mid \text{data})$, we need the posterior predictive distributions for a new π_* and for a new pair (γ_*, θ_*). Let $\boldsymbol{\phi}$ be the full parameter vector corresponding to model (8), that is, $\boldsymbol{\phi} = \{\pi_1, ..., \pi_{N_g}, (\gamma_1, \theta_1), ..., (\gamma_{N_g}, \theta_{N_g}),\ \beta, \delta, \mu_\pi\}$.

To obtain the expressions for $p(\pi_* \mid \text{data})$, $p((\gamma_*, \theta_*) \mid \text{data})$ and $p(y_* \mid \text{data})$, we need an expression for $p(y_*, \pi_*, (\gamma_*, \theta_*), \boldsymbol{\phi} \mid \text{data})$. This can be found by adding y_* to the first stage of model (6) and π_* and (γ_*, θ_*) to the second and third stages of model (6), and then again marginalizing G_1 and G_2 over their DP priors. Specifically,

$$\begin{aligned} p(y_*, \pi_*, (\gamma_*, \theta_*), \boldsymbol{\phi} \mid \text{data}) \ &= \ \{\pi_* I(y_* = 0) + (1 - \pi_*) \\ &\quad \times f(y_*; \gamma_*, \theta_*) I(y_* > 0)\} \\ &\quad \times p((\gamma_*, \theta_*) \mid (\gamma_1, \theta_1), ..., (\gamma_{N_g}, \theta_{N_g}), \beta, \delta) \\ &\quad \times p(\pi_* \mid \pi_1, ..., \pi_{N_g}, \mu_\pi) \times p(\boldsymbol{\phi} \mid \text{data}), \end{aligned} \tag{13}$$

where

$$p(\pi_* \mid \pi_1, ..., \pi_{N_g}, \mu_\pi) = \frac{\alpha_1}{\alpha_1 + N_g} g_{10}(\pi_*; \mu_\pi, \sigma_\pi^2) + \frac{1}{\alpha_1 + N_g} \sum_{i=1}^{N_g} \delta_{\pi_i}(\pi_*) \tag{14}$$

and

$$\begin{aligned} p((\gamma_*, \theta_*) \mid (\gamma_1, \theta_1), ..., (\gamma_{N_g}, \theta_{N_g}), \beta, \delta) = \frac{\alpha_2}{\alpha_2 + N_g} g_{20}((\gamma_*, \theta_*); \beta, \delta) + \\ \frac{1}{\alpha_2 + N_g} \sum_{i=1}^{N_g} \delta_{(\gamma_i, \theta_i)}(\gamma_*, \theta_*). \end{aligned} \tag{15}$$

Now, using the posterior samples for $\boldsymbol{\phi}$ (resulting from the MCMC algorithm described in Appendix B) and with appropriate integrations in expression (13), we can obtain posterior predictive inference for π_*, (γ_*, θ_*), and y_*. In particular,

$$p(\pi_* \mid \text{data}) = \int p(\pi_* \mid \pi_1, ..., \pi_{N_g}, \mu_\pi) p(\boldsymbol{\phi} \mid \text{data}) \mathrm{d}\boldsymbol{\phi}$$

and therefore posterior predictive draws for π_* can be obtained by drawing from (14) for each posterior sample for $\pi_1, ..., \pi_{N_g}, \mu_\pi$. Moreover,

$$p((\gamma_*, \theta_*) \mid \text{data}) = \int p((\gamma_*, \theta_*) \mid (\gamma_1, \theta_1), ..., (\gamma_{N_g}, \theta_{N_g}), \beta, \delta) p(\boldsymbol{\phi} \mid \text{data}) \mathrm{d}\boldsymbol{\phi}$$

can be sampled by drawing from (15) for each posterior sample for (γ_1, θ_1), ..., $(\gamma_{N_g}, \theta_{N_g}), \beta, \delta$. Finally,

$$\begin{aligned} p(y_* \mid \text{data}) \quad = \quad & \textstyle\int\int\int \{\pi_* I(y_* = 0) + (1 - \pi_*) f(y_*; \gamma_*, \theta_*) I(y_* > 0)\} \\ & \times p(\pi_* \mid \pi_1, ..., \pi_{N_g}, \mu_\pi) \\ & \times p((\gamma_*, \theta_*) \mid (\gamma_1, \theta_1), ..., (\gamma_{N_g}, \theta_{N_g}), \beta, \delta) \\ & \times p(\boldsymbol{\phi} \mid \text{data}) \; \mathrm{d}\pi_* \, \mathrm{d}(\gamma_*, \theta_*) \, \mathrm{d}\boldsymbol{\phi}. \end{aligned}$$

Based on this expression, posterior predictive samples for y_* can be obtained by first drawing π_* and (γ_*, θ_*) from the appropriate prior distributions based on the Dirichlet process functional (using expressions (14) and (15), respectively, for each posterior sample for $\boldsymbol{\phi}$) and then drawing y_* from $\pi_* I(y_* = 0) + (1 - \pi_*) f(y_*; \gamma_*, \theta_*) I(y_* > 0)$. Therefore, the posterior predictive distribution for a new group will have a point mass at 0 (driven by the posterior draws for π_*) and a continuous component (driven by the posterior draws for (γ_*, θ_*)).

Expressions (14) and (15) highlight the clustering structure induced by the DP priors, which enables flexible data-driven shapes in the posterior predictive densities $p(\pi_* \mid \text{data})$ and $p(\gamma_*, \theta_* \mid \text{data})$, and thus flexible tail behavior for the continuous component of $p(y_* \mid \text{data})$. The utility of such flexibility in the prior is illustrated in the following sections with both the simulated and the real data.

3. The simulation example

We now present a small simulation study to demonstrate the utility of the nonparametric approach. We simulated data for two cases; one case drew random-effects parameters from unimodal distributions, and one case drew random-effects parameters from multimodal distributions. We focus on prediction of the response of individuals in new groups because this is the setting where the nonparametric model offers the most promise.

All the simulated data were produced by first generating a $(\gamma_i, \theta_i, \pi_i)$ triple from the distributions we will outline. Then, using these parameters, data were generated for 100 groups with 30 observations in each group. The data were then analyzed using both the parametric and the nonparametric models.

In Case I (the unimodal case), the γ_i's were drawn from a Gamma$(2, 5)$, the θ_i's from a Gamma$(2, 10)$, and the π_i's from a Beta$(4, 5)$. The draws were independent, and given these parameters, the data were drawn according to the likelihood in (2).

In Case II (the multimodal case), the γ_i's were drawn from either a Gamma$(2, 1)$ or a Gamma$(50, 1)$ with equal probability. The θ_i's were drawn independently using the same scenario as the γ_i's, and the π_i's were drawn independently from either a Beta$(20, 80)$ or a Beta$(80, 20)$ with equal probability. Again, once the parameters were drawn, the data were produced using the likelihood in (2).

The parametric model was fitted using $\sigma_\pi^2 = 0.03$, $b = d = 1$, and $A_\beta = A_\delta = 40$, although sensitivity analyses showed that posterior distributions were virtually the same with other values of these parameters. These same values were used for the centering distributions of the nonparametric model. Also, we chose to use $\alpha_1 = \alpha_2 = 2$ to analyze simulation data. We used $50,000$ burn-in iterations for both models. We followed the burn-in with $100,000$ posterior draws, keeping every 10^{th} draw for the parametric model, and with $1,000,000$ posterior draws, keeping every 100^{th} draw for the nonparametric model. The nonparametric model results in higher correlation among posterior draws, and the higher thinning rate assures that the draws have converged appropriately to the posterior distribution.

Now we review the reason behind our simulation choices. In Case I, the parametric priors we have previously described are correct and should yield appropriate prediction. In Case II, the parametric priors are not suitable, so one might expect prediction to be problematic. However, we used the same nonparametric model in both cases. That is, we let the DP prior structure identify the appropriate form in both cases. If the nonparametric formulation is successful, it won't matter what the true prior is, since the nonparametric model will find it.

The simulation results convey two main messages. The first message is that the parametric model will not replicate the modes unless they are an explicit part of the prior formulation when predicting parameters for new groups, while the nonparametric methodology performs this task quite well because the modes do not need to be an explicit part of the prior formulation. Figures 1 and 2 demonstrate this. In Figure 1, we see the results from Case I, the unimodal case. The posterior densities from the parametric model follow the generated parameter histograms quite closely. The nonparametric model produces comparable results. However, in Figure 2, it is obvious that the parametric model cannot predict the multiple modes. The nonparametric model does this quite well since the prior distributions are covered by the functional forms supported by the DP priors. This means that unless the possibility of multiple modes is explicitly addressed in the parametric setting (a practically impossible task if only data are examined since the multimodality occurs in the distributions of the parameters and not in the distributions of the data itself), it would be unreasonable to expect the parametric model to predict efficiently. On the other hand, the nonparametric model will automatically handle the problem with absolutely no change in the code.

The second message is that the posterior point estimation of parameters for the groups represented in the simulated data sets is quite similar for both models. In Figures 3, 4, and 5, we show posterior intervals (5th to 95th percentiles) for each group in simulation Case II. Remember, in Case II the parametric priors are not suitable. Nonetheless, it is clear that both methods separate the modes in the prior densities quite well for the estimated parameters. It is interesting that the posterior intervals are generally wider for the parametric model. This greater width may be explained by examining Figure 2. Since the parametric model must span the space of the multiple modes with only a single peak, much of the distribution is over space where no parameters occur. Thus, uncertainty regarding the location of the parameters is overestimated. Misspecification of the prior can lead to artificially high uncertainty regarding the parameter estimates.

4. The data

The data set is taken from a major medical plan, covering a block of medium-sized groups in Illinois and Wisconsin for 1994 and 1995. Each policyholder was part of a group plan. In 1994 the groups consisted of 1 to 103 employees with a median size of 5 and an average size of 8.3. We have claims information on 8,921 policyholders from 1,075 groups. Policies were all of the same type (employee plus one individual). Table 1 gives some descriptive summary information about the data in both 1994 and 1995.

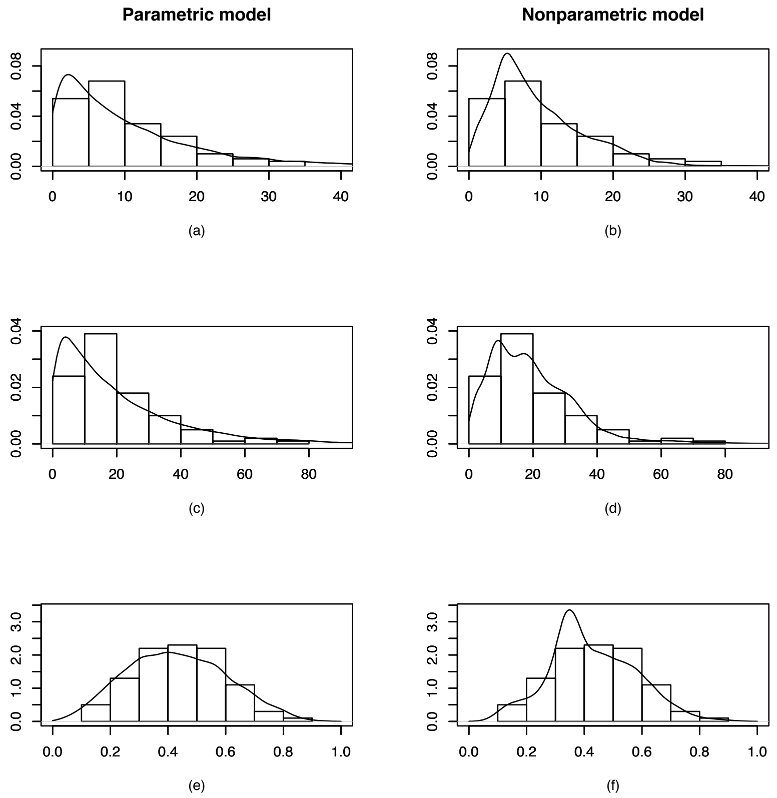

Figure 1: Simulation Case I—unimodal priors. Posterior densities for γ_* (panels (a) and (b)), for θ_* (panels (c) and (d)), and for π_* (panels (e) and (f)), under the parametric model (left column) and the nonparametric model (right column). The histograms plot the generated γ_i (panels (a) and (b)), θ_i (panels (c) and (d)), and π_i (panels (e) and (f)), $i = 1, ..., 100$.

Table 1: Descriptive information for the data analyzed from both 1994 and 1995.

	n obs.	n groups	Mean	Std. Dev.	Median	Maximum	Percentage Zero Claims
1994	8921	1075	6.79	21.01	1.11	643.02	.315
1995	8732	1129	5.18	11.63	0.88	297.30	.357

Although the data are dated from a business perspective, they provide

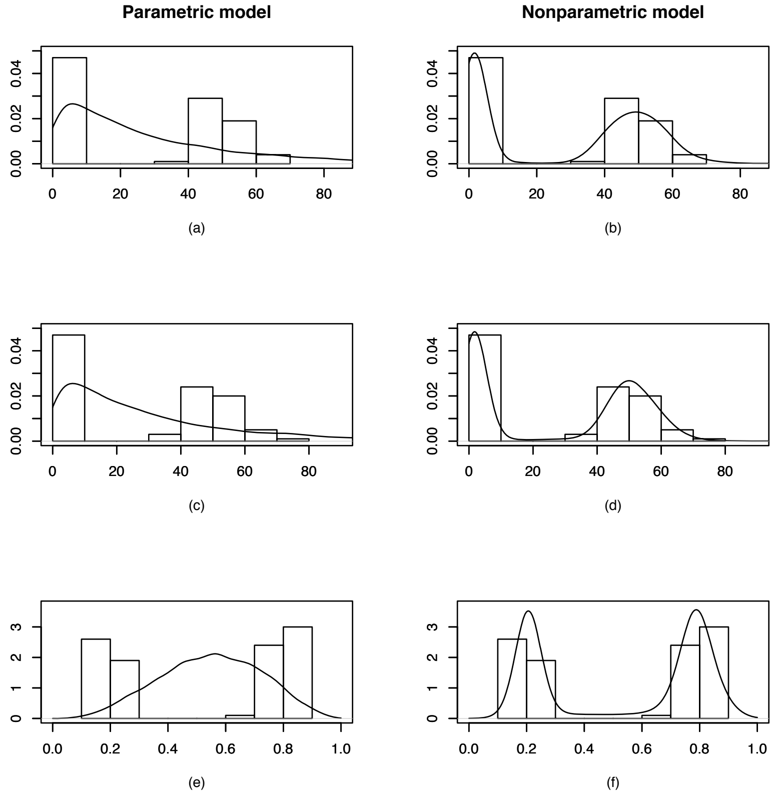

Figure 2: Simulation Case II—multimodal priors. Posterior densities for γ_* (panels (a) and (b)), for θ_* (panels (c) and (d)), and for π_* (panels (e) and (f)), under the parametric model (left column) and the nonparametric model (right column). The histograms plot the generated γ_i (panels (a) and (b)), θ_i (panels (c) and (d)), and π_i (panels (e) and (f)), $i = 1, ..., 100$.

an opportunity to compare the parametric and nonparametric paradigms without divulging proprietary information.

Total costs, including deductible and copayments, were accrued by each policyholder on a yearly basis. The total yearly costs were then divided by the number of days the policy was in force during the year. As per the policy of the company providing the data, all policies with annual claims costs exceeding $25,000 were excluded from all analyses. Large daily costs are still possible if the policy was in force for only a small number of days but is associated with relatively large total costs.

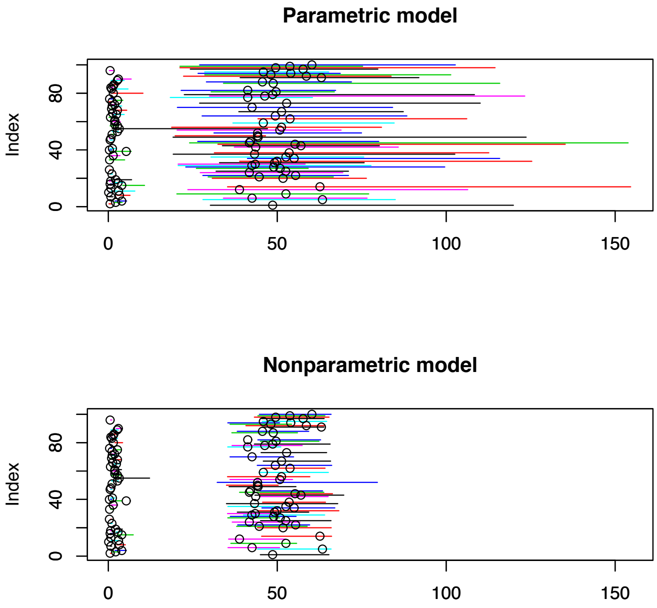

Figure 3: Simulation Case II. Posterior intervals (5th to 95th posterior percentile) for each γ_i, $i = 1, ..., 100$ under the parametric (upper panel) and nonparametric (lower panel) models. The circles denote the actual generated γ_i.

5. Analysis of the claims data

The 1994 data consists of $8,921$ observations in $1,075$ groups. Because of work with other data of the same type, we expected the γ_i with the actual data to be smaller than the γ_i we used when we simulated data. Thus, we used $A_\beta = 3$, while A_δ remained relatively large at 30 in both the parametric and nonparametric settings. For the data analysis we used $\alpha_1 = \alpha_2 = 3$. In both models we used a burn-in of $50,000$ with $100,000$ posterior draws, keeping every 10th draw. Both models displayed convergent chains for the posterior draws of all parameter densities (Raftery and Lewis, 1996; Smith, 2005).

In Figure 6, we show posterior densities for both the parametric and nonparametric models for the γ_*, θ_*, and π_*. We note that the nonparametric model posterior densities showed multimodal behavior like those demonstrated in Case II of the simulation study. This multimodal behavior would be virtually impossible to uncover prior to the analysis since it is in the dis-

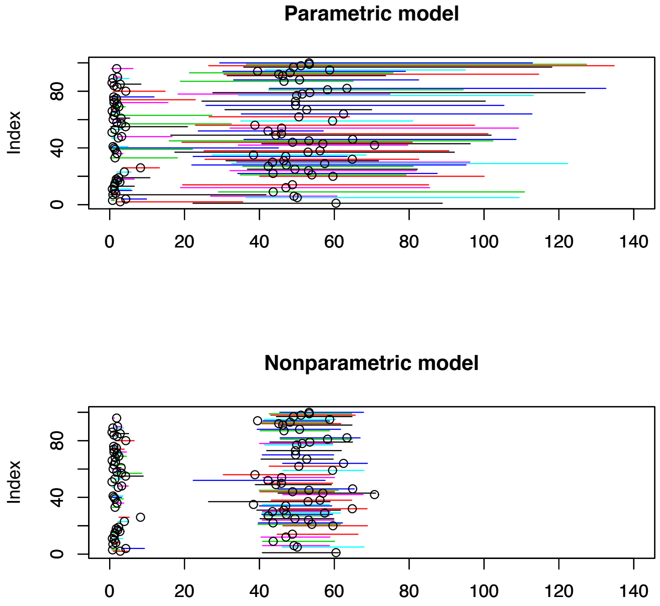

Figure 4: Simulation Case II. Posterior intervals (5^{th} to 95^{th} posterior percentile) for each $\theta_i,\ i = 1, ..., 100$ under the parametric (upper panel) and nonparametric (lower panel) models. The circles denote the actual generated θ_i.

tributions of the parameters, not the distribution of the data. Use of the DP prior offers a flexible way to uncover such nonstandard distributional shapes.

Since the densities actually have this multimodality, we anticipate that the nonparametric model will do better in predicting costs from new groups. We would, however, expect that predicting behavior in groups already present in the data would be quite similar for the two approaches, as was displayed in the simulation. Also, we would not be surprised by an overestimation of uncertainty in the parameter estimates under the parametric model. Again, we emphasize that there is no way to uncover this kind of multimodality in the parameters without using a methodology that spans this kind of behavior in the prior specifications. There is no way to anticipate this kind of structure solely by examining the data.

We reemphasize at this point why the prior distributions of the parameters are of such interest when we are predicting values for costs. Predicting new costs depends on drawing reasonable new values of the parameters. Since

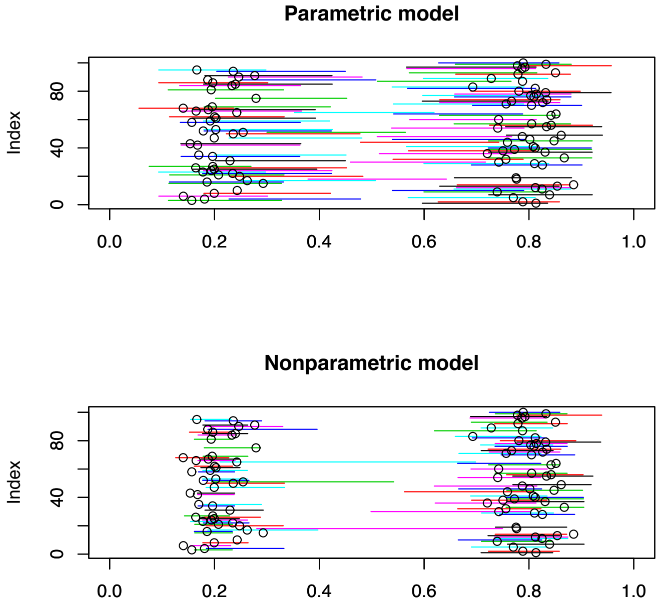

Figure 5: Simulation Case II. Posterior intervals (5^{th} to 95^{th} posterior percentile) for each π_i, $i = 1, ..., 100$ under the parametric (upper panel) and nonparametric (lower panel) models. The circles denote the actual generated π_i.

the predictive distributions of the parameters are based on the prior specification of the parameters, it is imperative that these prior specifications be flexible if we are going to get accurate predictions of new data.

We chose one group that had fairly large representation in both 1994 and 1995 to check the assertion that both methods should be quite similar in predicting behavior for a group already present in the data. Group 69511 had 81 members in 1994 and 72 members in 1995. We had no way to determine how many members were the same in both years. Using posterior samples from the corresponding triple $(\pi_i, \gamma_i, \theta_i)$, we obtained the posterior predictive distribution for this group using both models. In Figure 7 (left panel), we show the posterior predictive distribution for the nonzero data for both the parametric and the nonparametric model as well as the histogram of the actual 1995 nonzero data for that group. There is little difference in the posterior predictive distributions, and both distributions model the 1995 data reasonably well.

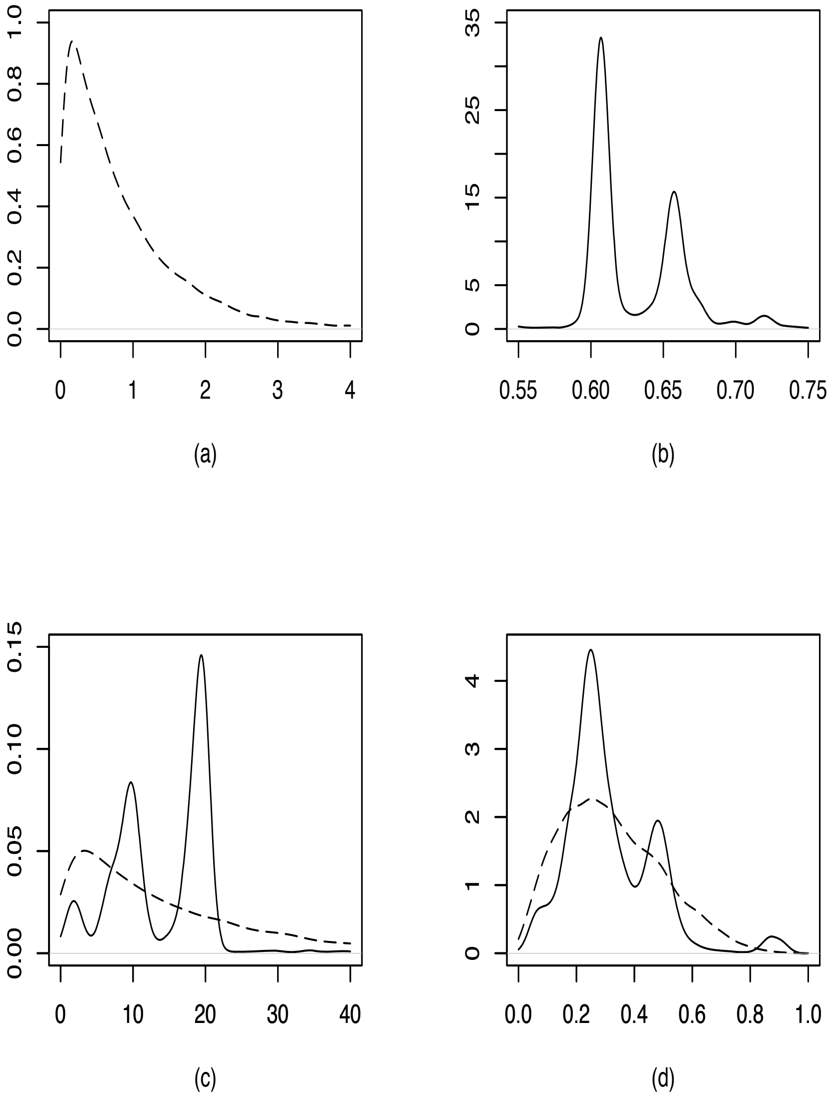

Figure 6: Posterior predictive inference for the random-effects distributions for the real data. Panels (a) and (b) include the posterior density for γ_* under the parametric and nonparametric models, respectively. (Note the different scale in these two panels.) The posterior densities for θ_* and for π_* are shown in panels (c) and (d), respectively; in all cases, the solid lines correspond to the nonparametric model and the dashed lines correspond to the parametric model.

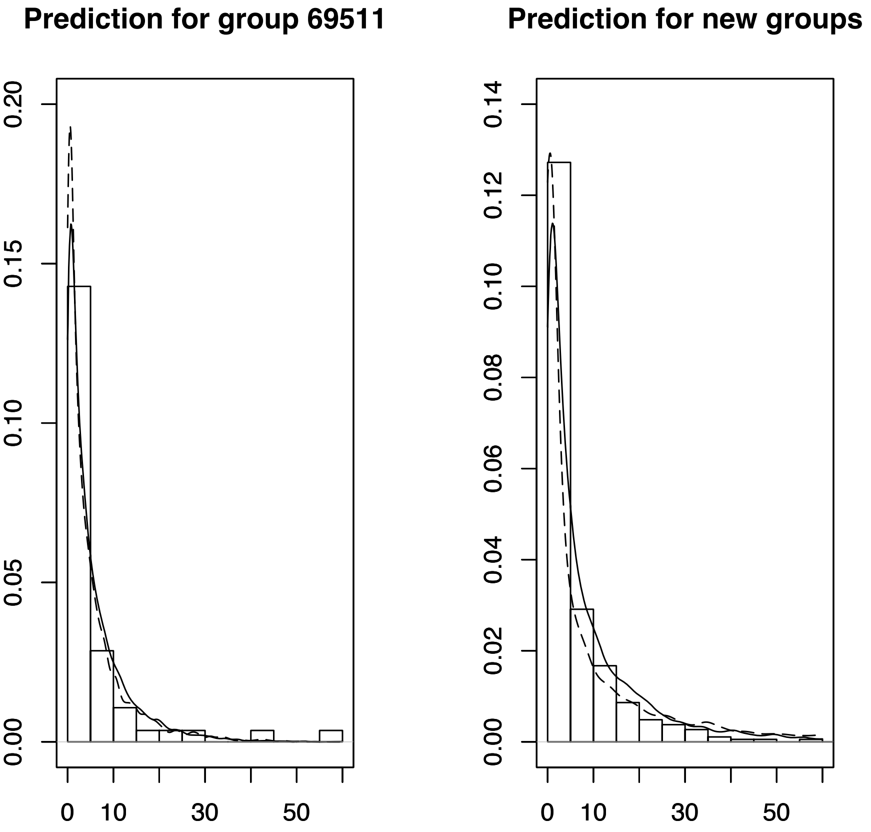

Figure 7: Cross-validated posterior predictive inference for the real data. Posterior results are based on data from year 1994 and are validated using corresponding data from year 1995 (given by the histograms in the two panels). The left panel includes posterior predictive densities for claims under group 69511. Posterior predictive densities for claims under a new group are plotted on the right panel. In both panels, solid and dashed lines correspond to the nonparametric model and parametric model, respectively.

To further quantify the differences between the two models, we computed a model comparison criterion that focuses on posterior predictive inference. If y_{0j}, $j = 1, \ldots, J$, represent the non-zero observations from group 69511 in 1995, we can estimate $p(y_{0j} \mid \text{data})$, i.e., the conditional predictive ordinate (CPO) at y_{0j}, using $B^{-1} \sum_{b=1}^{B} f(y_{0j}; \gamma_{*,b}, \theta_{*,b})$, where $\{(\gamma_{*,b}, \theta_{*,b}) : b = 1, \ldots, B\}$ is the posterior predictive sample for (γ_*, θ_*) ($B = 10,000$ in our analysis). Note that these are cross-validation posterior predictive calculations, since the 1995 data y_{0j} were not used in obtaining the posterior distribution for the model. We expect the CPO for a given data point to be higher in the model that has a better predictive fit. Of the $J = 56$ non-zero observations in 1995, 47 CPO values were greater for the nonparametric model (84%). The CPO values can also be summarized using the cross-validation posterior predictive criterion given by $Q = J^{-1} \sum_{j=1}^{J} \log(p(y_{0j} \mid \text{data}))$ (e.g., Bernardo and Smith, 2009). A bigger value of Q implies more predictive ability. For the parametric model, we obtain $Q = -2.86$, while for the nonparametric model $Q = -2.60$. Thus, the predictive ability of the nonparametric model exceeded that of the parametric model for these data.

Next, we focused on predicting outcomes in 1995 for groups not present in the 1994 data. There were $8,732$ observations in 1995, and 522 of these observations came from 101 groups that were not represented in 1994. We treated these 522 observations as if they came from one new group and estimated posterior predictive densities for this new group under both the parametric and nonparametric models. In Figure 7 (right panel), we show the posterior predictive densities for positive claim costs from a new group overlaid on the histogram of the corresponding 1995 data. Here, we observe that the posterior predictive distributions of the two models differ, with the nonparametric model having a higher density over the mid-range of the responses than the parametric model.

Of the $J = 371$ non-zero observations in 1995, 327 CPO values were greater for the nonparametric model (88%). For the parametric model, we obtain $Q = -3.20$, while for the nonparametric model $Q = -2.94$. Thus, the predictive ability of the nonparametric model exceeded that of the parametric model both for a group present in both data sets, and for new groups not present in the 1994 data.

6. Discussion

Bayesian nonparametric methods provide a class of models that offer substantial advantages in predictive modeling. They place prior distributions on spaces of functions rather than placing the distributions on parameters of a specific function. This broadening of the prior space allows for priors that may have quite different properties (e.g., multiple modes) than those anticipated.

In the data we examined, the presence of multiple modes in the predictive distributions for the parameters was not anticipated. However, a posteriori we can postulate an explanation. If we think of the general population as being relatively healthy, then we would expect most groups to reflect this state. However, if there are a few individuals in some groups with less-than-perfect health (i.e., more frail), we would expect to see longer tails in these groups. Some small proportion of the groups might have extremely long tails. Figure 6 illustrates this pattern. The lowest mode of the posterior distribution of the γ_i's is generally associated with the largest mode of the θ_i's. That is, groups with γ_i in a range of 0.59 to 0.63 tend to be associated with θ_i in the range of 13 to 20. In fact, the mean of the θ_i's associated with γ_i's in the range of 0.59 to 0.63 is 18.5. Also, the middle modes of the two distributions tend to be associated (the mean of the θ_i's associated with γ_i's in the range of 0.65 to 0.68 is 13.6) and the highest mode of the γ_i's tends to go with the smallest mode of the θ_i's. Since these distributions are parameterized to have means of $\gamma \times \theta$ and variances of $\gamma \times \theta^2$, we see that the means of the groups are relatively stable, while the variances for some groups are quite a bit larger. This type of cost experience might be due to the age of the clients, but other explanations are equally plausible. It might just as well result from serious illness associated with one or two members of relatively small numbers of groups. So it is possible, though unlikely, that the parametric model might be able to perform on a par with the nonparametric model with a complete inclusion of possible covariates in the model. The problem, of course, is that failing to measure important covariates is a common and ongoing issue in predictive modeling.

While the association between frailty and the multimodal behavior of the distributions of the parameters may seem reasonable in retrospect, it would not be obvious before completing the analysis, and it would not be uncovered at all using a conventional parametric analysis. Thus, a procedure that allows for great flexibility in the specification of prior distributions can pay large dividends.

Bayesian nonparametric modeling offers high utility to the practicing actuary as it allows for prediction that cannot be matched by the traditional Bayesian approach. This added ability to predict costs with greater accuracy will improve risk management.

Appendix A. Dirichlet process priors and Dirichlet process mixtures

Here, we provide a brief review of Dirichlet processes (DPs) and DP mixture models. The main theoretical results on inference for DP mixtures can be found in the work of Antoniak (1974). For early work on modeling and inference using DP mixtures see, for example,Brunner and Lo (1989); Ferguson (1983); Kuo (1986); Lo (1984).

Appendix A.1. The Dirichlet process

The Dirichlet Process (Ferguson, 1973, 1974) is a stochastic process with sample paths that can be interpreted as distributions G (equivalently, CDFs) on a sample space Ω. The DP can be defined in terms of two parameters: a positive scalar parameter α, which can be interpreted as a precision parameter, and a specified base (centering) distribution G_0 on Ω. For example, when $\Omega = R$, for any $x \in R$, $G(x)$ has a Beta distribution with parameters $\alpha\, G_0(x)$ and $\alpha[1 - G_0(x)]$; thus, $\mathrm{E}[G(x)] = G_0(x)$ and $\mathrm{Var}[G(x)] = G_0(x)[1-G_0(x)]/(\alpha+1)$. Hence, for larger values of α, a realization G from the DP is expected to be closer to the base distribution G_0. We write $G \sim \mathrm{DP}(\alpha, G_0)$ to denote that a DP prior is used for the random CDF (distribution) G. In fact, DP-based modeling typically utilizes mixtures of DPs (Antoniak, 1974); that is, a more flexible version of the DP prior that involves hyperpriors for α or the parameters $\boldsymbol{\psi}$ of $G_0(\cdot) \equiv G_0(\cdot|\boldsymbol{\psi})$, or both.

A practical, useful definition of the DP was given by Sethuraman (1994). According to this constructive definition, a realization G from $\mathrm{DP}(\alpha, G_0)$ is (almost surely) of the form

$$G(\cdot) = \sum_{i=1}^{\infty} w_i\, \delta_{\vartheta_i}(\cdot),$$

where $\delta_x(\cdot)$ denotes a point mass at x. Here, the ϑ_j are i.i.d. G_0, and the weights are constructed through a *stick-breaking* procedure $w_1 = z_1$, $w_i = z_i \prod_{k=1}^{i-1}(1 - z_k), i = 2, 3, \ldots$, with the z_k i.i.d. $\text{Beta}(1, \alpha)$; moreover, the sequences $\{z_k, k = 1, 2, \ldots\}$ and $\{\vartheta_j, j = 1, 2, \ldots\}$ are independent. Hence, the DP generates, with probability one, discrete distributions that can be represented as countable mixtures of point masses, with locations drawn independently from G_0 and weights generated according to a stick-breaking mechanism based on i.i.d. draws from a $\text{Beta}(1, \alpha)$ distribution.

The DP constructive definition has motivated extensions of the DP in several directions, including priors with more general structure (e.g., Ishwaran and James, 2001) and prior models for dependent distributions (e.g., De Iorio et al., 2004; Gelfand et al., 2005; Griffin and Steel, 2006).

Appendix A.2. Dirichlet process mixture models

A natural way to increase the applicability of DP-based modeling is by using the DP as a prior for the mixing distribution in a mixture model with a parametric kernel distribution $K(\cdot|\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta \subseteq R^p$ (with corresponding *density* — probability density or probability mass function — $k(\cdot|\boldsymbol{\theta})$). This approach yields the class of DP mixture models, which can be generically expressed as

$$F(\cdot; G) = \int K(\cdot|\boldsymbol{\theta}) \, \mathrm{d}G(\boldsymbol{\theta}), \qquad G \mid \alpha, \boldsymbol{\psi} \sim \text{DP}(\alpha, G_0(\cdot|\boldsymbol{\psi})),$$

with the analogous notation for the random mixture density, $f(\cdot; G) = \int k(\cdot|\boldsymbol{\theta}) \, \mathrm{d}G(\boldsymbol{\theta})$. The kernel can be chosen to be a possibly multivariate continuous distribution, thus overcoming the almost sure discreteness of the DP.

Consider $F(\cdot; G)$ as the model for the stochastic mechanism corresponding to data $\boldsymbol{Y} = (Y_1, ..., Y_n)$. For example, assume Y_i, given G, i.i.d. from $F(\cdot; G)$ with the DP prior structure for G. Working with this generic DP mixture model typically involves the introduction of a vector of latent mixing parameters, $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_n)$, where $\boldsymbol{\theta}_i$ is associated with Y_i, such that the model can be expressed in hierarchical form as follows:

$$\begin{array}{rcc} Y_i|\boldsymbol{\theta}_i & \overset{\text{ind.}}{\sim} & K(\cdot|\boldsymbol{\theta}_i), \;\; i = 1, ..., n \\ \boldsymbol{\theta}_i|G & \overset{\text{i.i.d.}}{\sim} & G, \;\; i = 1, ..., n \\ G \mid \alpha, \boldsymbol{\psi} & \sim & \text{DP}(\alpha, G_0(\cdot|\boldsymbol{\psi})). \end{array} \tag{A.1}$$

The model can be completed with priors for α and $\boldsymbol{\psi}$. Moreover, practically important semiparametric versions can be developed by working with kernels $K(\cdot|\boldsymbol{\theta}, \boldsymbol{\phi})$ where the $\boldsymbol{\phi}$ portion of the parameter vector is modelled parametrically; for example, $\boldsymbol{\phi}$ could be a vector of regression coefficients incorporating a regression component in the model.

The Pólya urn DP characterization (Blackwell and MacQueen, 1973) is key in the DP mixture setting, since it results in a practically useful version of (A.1) where G is marginalized over its DP prior. The resulting joint prior for the $\boldsymbol{\theta}_i$ is given by

$$p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n|\alpha, \boldsymbol{\psi}) = G_0(\boldsymbol{\theta}_1) \prod_{i=2}^{n} \left\{ \frac{\alpha}{\alpha + i - 1} G_0(\boldsymbol{\theta}_i) + \frac{1}{\alpha + i - 1} \sum_{\ell=1}^{i-1} \delta_{\boldsymbol{\theta}_\ell}(\boldsymbol{\theta}_i) \right\}.$$

This result is central to the development of posterior simulation methods for DP mixtures (see, for example, the reviews in Müller and Quintana, 2004; Hanson et al., 2005).

This class of Bayesian nonparametric models is now arguably the most widely used due to the availability of several posterior simulation techniques based on MCMC algorithms. For example, see Bush and MacEachern (1996); Escobar and West (1995); Gelfand and Kottas (2002); Ishwaran and James (2001); Jain and Neal (2004); MacEachern (1998); Neal (2000). See Blei and Jordan (2006); Liu (1996); MacEachern et al. (1999) for alternative approaches.

Appendix B. The MCMC algorithm for the nonparametric model

The joint posterior, $p(\pi_1, ..., \pi_{N_g}, (\gamma_1, \theta_1), ..., (\gamma_{N_g}, \theta_{N_g}), \beta, \delta, \mu_\pi \mid \text{data})$, corresponding to model (8) is proportional to

$$p(\beta)p(\delta)p(\mu_\pi)p(\pi_1, ..., \pi_{N_g} \mid \mu_\pi)p((\gamma_1, \theta_1), ..., (\gamma_{N_g}, \theta_{N_g}) \mid \beta, \delta) \\ \times \left\{ \prod_{i=1}^{N_g} \pi_i^{L_{i0}} (1 - \pi_i)^{L_i - L_{i0}} \right\} \left\{ \prod_{i=1}^{N_g} \prod_{\{\ell : y_{i\ell} > 0\}} f(y_{i\ell}; \gamma_i, \theta_i) \right\}, \quad \text{(B.1)}$$

where $L_{i0} = |\{\ell : y_{i\ell} = 0\}|$, so that $|\{\ell : y_{i\ell} > 0\}| = L_i - L_{i0}$.

The MCMC algorithm involves Metropolis-Hastings (M-H) updates for each of the π_i and for each pair (γ_i, θ_i) using the prior full conditionals in (11) and (12) as proposal distributions. Updates are also needed for β, δ, and μ_π. Details on the steps of the MCMC algorithm are provided below.

1. Updating the π_i: For each $i = 1, ..., N_g$, the posterior full conditional for π_i is given by

$$p(\pi_i \mid ..., \text{data}) \propto p(\pi_i \mid \{\pi_j : j \neq i\}, \mu_\pi) \times \pi_i^{L_{i0}} (1 - \pi_i)^{L_i - L_{i0}},$$

   with $p(\pi_i \mid \{\pi_j : j \neq i\}, \mu_\pi)$ defined in (11). We use the following M-H update:

   - Let $\pi_i^{(\text{old})}$ be the current state of the chain. Repeat the following update R_1 times ($R_1 \geq 1$).
   - Draw a candidate $\tilde{\pi}_i$ from $p(\pi_i \mid \{\pi_j : j \neq i\}, \mu_\pi)$ using the form in equation (11).
   - Set $\pi_i = \tilde{\pi}_i$ with probability

$$q_1 = \min \left\{ 1, \frac{\tilde{\pi}_i^{L_{i0}} (1 - \tilde{\pi}_i)^{L_i - L_{i0}}}{\pi_i^{(\text{old}) L_{i0}} (1 - \pi_i^{(\text{old})})^{L_i - L_{i0}}} \right\},$$

     and $\pi_i = \pi_i^{(\text{old})}$ with probability $1 - q_1$.

2. Updating the (γ_i, θ_i): For each $i = 1, ..., N_g$, the posterior full conditional for (γ_i, θ_i) is

$$p((\gamma_i, \theta_i) \mid ..., \text{data}) \propto p((\gamma_i, \theta_i) \mid \{(\gamma_j, \theta_j) : j \neq i\}, \beta, \delta) \times \prod_{\{\ell : y_{i\ell} > 0\}} f(y_{i\ell}; \gamma_i, \theta_i), \quad \text{(B.2)}$$

   where $p((\gamma_i, \theta_i) \mid \{(\gamma_j, \theta_j) : j \neq i\}, \beta, \delta)$ is given by expression (12). The M-H step proceeds as follows:

   - Let $(\gamma_i^{(\text{old})}, \theta_i^{(\text{old})})$ be the current state of the chain. Repeat the following update R_2 times ($R_2 \geq 1$).
   - Draw a candidate $(\tilde{\gamma}_i, \tilde{\theta}_i)$ from distribution $p((\gamma_i, \theta_i) \mid \{(\gamma_j, \theta_j) : j \neq i\}, \beta, \delta)$ using the form in equation (12).

- Set $(\gamma_i, \theta_i) = (\tilde{\gamma}_i, \tilde{\theta}_i)$ with probability

$$q_2 = \min\left\{1, \frac{\prod_{\{\ell: y_{i\ell}>0\}} f(y_{i\ell}; \tilde{\gamma}_i, \tilde{\theta}_i)}{\prod_{\{\ell: y_{i\ell}>0\}} f(y_{i\ell}; \gamma_i^{(\text{old})}, \theta_i^{(\text{old})})}\right\},$$

  and $(\gamma_i, \theta_i) = (\gamma_i^{(\text{old})}, \theta_i^{(\text{old})})$ with probability $1 - q_2$.

3. Updating the hyperparameters: Once all the π_i, $i = 1, ..., N_g$ are updated, we obtain N_1^* $(\leq N_g)$, the number of distinct π_i, and the distinct values π_j^*, $j = 1, ..., N_1^*$. Similarly, after updating all the (γ_i, θ_i), $i = 1, ..., N_g$, we obtain a number N_2^* $(\leq N_g)$ of distinct (γ_i, θ_i) with distinct values (γ_j^*, θ_j^*), $j = 1, ..., N_2^*$.
   Now, the posterior full conditional for β can be expressed as

$$p(\beta \mid ..., \text{data}) \propto \beta^{-3} \exp(-A_\beta/\beta) \times \prod_{j=1}^{N_2^*} \text{Gamma}(\gamma_j^*; b, \beta),$$

   so

$$\begin{aligned} p(\beta \mid ..., \text{data}) &\propto \beta^{-3} \exp(-A_\beta/\beta) \times \prod_{j=1}^{N_2^*} \beta^{-b} \exp(-\gamma_j^*/\beta) \\ &\propto \beta^{-(bN_2^*+3)} \exp(-(A_\beta + \textstyle\sum_{j=1}^{N_2^*} \gamma_j^*)/\beta); \end{aligned} \quad \text{(B.3)}$$

   therefore, we recognize the posterior full conditional for β as an inverse gamma distribution with shape parameter $bN_2^* + 2$ and scale parameter $A_\beta + \sum_{j=1}^{N_2^*} \gamma_j^*$.
   Analogously, the posterior full conditional for δ is

$$p(\delta \mid ..., \text{data}) \propto \delta^{-3} \exp(-A_\delta/\delta) \times \prod_{j=1}^{N_2^*} \text{gamma}(\theta_j^*; d, \delta),$$

   and we therefore obtain an inverse gamma posterior full conditional distribution for δ with shape parameter $dN_2^* + 2$ and scale parameter $A_\delta + \sum_{j=1}^{N_2^*} \theta_j^*$.

Finally, the posterior full conditional for μ_π is given by

$$p(\mu_\pi \mid ..., \text{data}) \propto p(\mu_\pi) \times \prod_{j=1}^{N_1^*} g_{10}(\pi_j^*; \mu_\pi, \sigma_\pi^2),$$

and this does not lead to a distributional form that can be sampled directly. An M-H step was used with a normal proposal distribution centered at the current state of the chain and tuned with the variance to achieve an appropriate acceptance rate.

References

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. The Annals of Statistics 2(6), 1152–1174.

Bernardo, J. M. and A. F. Smith (2009). Bayesian Theory, Volume 405. Wiley.

Blackwell, D. and J. MacQueen (1973). Ferguson distributions via Pólya urn schemes. The Annals of Statistics 1(2), 353–355.

Blei, D. M. and M. I. Jordan (2006). Variational inference for Dirichlet process mixtures. Bayesian Analysis 1(1), 121–143.

Box, G. E. and N. R. Draper (1987). Empirical Model-building and Response Surfaces: Wiley Series in Probability and Mathematical Statistics. John Willey & Sons New York.

Brunner, L. J. and A. Y. Lo (1989). Bayes methods for a symmetric unimodal density and its mode. The Annals of Statistics 17(4), 1550–1566.

Bush, C. A. and S. N. MacEachern (1996). A semiparametric Bayesian model for randomised block designs. Biometrika 83(2), 275–285.

De Iorio, M., P. Müller, G. L. Rosner, and S. N. MacEachern (2004). An ANOVA model for dependent random measures. Journal of the American Statistical Association 99(465), 205–215.

Dey, D., P. Müller, and D. Sinha (1998). Practical Nonparametric and Semiparametric Bayesian Statistics. Springer Heidelberg.

Escobar, M. D. and M. West (1995, Jun). Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association 90(430), 577–588.

Fellingham, G. W., H. Dennis Tolley, and T. N. Herzog (2005). Comparing credibility estimates of health insurance claims costs. North American Actuarial Journal 9(1), 1–12.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. The Annals of Statistics 1(2), 209–230.

Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics*, 615–629.

Ferguson, T. S. (1983). Bayesian density estimation by mixtures of normal distributions. *Recent Advances in Statistics* *24*, 287–302.

Forbes, C., M. Evans, N. Hastings, and B. Peacock (2011). *Statistical Distributions*. Wiley.

Gelfand, A. E. and A. Kottas (2002). A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* *11*(2).

Gelfand, A. E., A. Kottas, and S. N. MacEachern (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association* *100*(471), 1021–1035.

Gilks, W. R., N. G. Best, and K. K. C. Tan (1995). Adaptive rejection metropolis sampling within Gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* *44*(4).

Griffin, J. E. and M. J. Steel (2006). Order-based dependent Dirichlet processes. *Journal of the American statistical Association* *101*(473), 179–194.

Hanson, T., A. Branscum, and W. Johnson (2005). Bayesian nonparametric modeling and data analysis: an introduction. *Handbook of Statistics* *25*, 245–278.

Ishwaran, H. and L. F. James (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* *96*(453), 161–173.

Jain, S. and R. M. Neal (2004, Mar). A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics* *13*(1), 158–182.

Kuo, L. (1986). Computations of mixtures of Dirichlet processes. *SIAM Journal on Scientific and Statistical Computing* *7*(1), 60–71.

Liu, J. S. (1996). Nonparametric hierarchical Bayes via sequential imputations. *The Annals of Statistics* *24*(3), 911–930.

Lo, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. density estimates. The Annals of Statistics 12(1), 351–357.

MacEachern, S. N. (1998). Estimating normal means with a conjugate style Dirichlet process prior. In D. D. Dey, P. Müller, and D. Sinha (Eds.), Practical Nonparametric and Semiparametric Bayesian Statistics, pp. 23–43. New York: Springer-Verlag.

MacEachern, S. N., M. Clyde, and J. S. Liu (1999). Sequential importance sampling for nonparametric Bayes models: The next generation. Canadian Journal of Statistics 27(2), 251–267.

Müller, P. and F. Quintana (2004). Nonparametric Bayesian data analysis. Statistical Science 19(1), 95–110.

Neal, R. M. (2000, Jun). Markov chain sampling methods for Dirichlet process mixture models. Journal of Computational and Graphical Statistics 9(2), 249–265.

Raftery, A. E. and S. M. Lewis (1996). Implementing MCMC. Markov chain Monte Carlo in Practice, 115–130.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. Statistica Sinica 4(2), 639–650.

Smith, B. J. (2005). Bayesian output analysis program (boa). http://www.public-health.uiowa.edu/boa/.

Walker, S. G., P. Damien, P. W. Laud, and A. F. Smith (1999). Bayesian nonparametric inference for random distributions and related functions. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61(3), 485–527.