Parametric and Nonparametric Bayesian Methods to Model Health Insurance Claims Costs

Gilbert W. Fellingham, Brigham Young University
Athanasios Kottas, University of California, Santa Cruz

Abstract

We develop Bayesian parametric and nonparametric hierarchical approaches to modeling health insurance claims data. Both prediction methods produce credibility type estimators, which use relevant information from related experience. In the parametric model, the likelihood arises through a mixture of a gamma distribution for the non-zero costs (severity), with a point mass for the zero costs (propensity). In the nonparametric extension, Dirichlet process priors are associated with the propensity parameters as well as the severity parameters. Posterior inference and prediction for both models is based on Markov chain Monte Carlo posterior simulation methods. A simulation study is used to demonstrate the utility of the nonparametric model across different settings. Moreover, we illustrate the methodology using real data from 1994 and 1995 provided by a major medical provider from a block of medium sized groups in the Midwest. The models were fit to the 1994 data, with their performance assessed and compared using the 1995 data.

1 Introduction

The purpose of this paper is to introduce the practicing actuary to a flexible class of Bayesian models called Bayesian nonparametric models. Bayesian models provide a coherent way of incorporating prior information into the information contained in data, to produce an updated set of probability functions describing an individual's current uncertainty regarding the state of nature. For the actuary, tasks where the Bayesian paradigm makes particular sense are those involving modeling costs with an eye to predicting expected costs for the coming year. These models could be used in premium calculations for small groups, and in premium calculations for blocks of business in new areas, as well as to calculate experience based refunds.

In the Bayesian framework, the model consists of the likelihood of the data given the parameters, multiplied by probability densities for each of the parameters. The densities on the parameters are called the "prior" probabilities as they are formulated prior to the collection of the data. Based on Bayes theorem, posterior densities for the parameters given the data are then available from the scaled product of the likelihood and the priors. (For a review of Bayesian methods in general see e.g. Gelman, et al, 1998, Klugman, 1992, Scollnik, 2001, or Makov, 2001.) Thinking somewhat simplistically, Bayesian model specification hinges on selecting scientifically appropriate prior distributions.

If $(y_1, y_2, ..., y_n)$ represents the data, de Finetti (1937) showed that if $(y_1, ..., y_n)$ is part of an infinitely exchangeable sequence, all coherent joint predictive distributions $p(y_1, ..., y_n)$

must have the hierarchical form

$$\begin{array}{rcl} F & \sim & p(F) \\ (y_i \mid F) & \overset{\text{i.i.d.}}{\sim} & F, \end{array} \tag{1}$$

where F is the limiting empirical cumulative distribution function (CDF) of the infinite sequence $(y_1, y_2, ...)$. Thus, the Bayesian model specification task becomes choosing the scientifically appropriate prior distribution $p(F)$ for F. However, since F is an infinite dimensional parameter, putting the appropriate probability distribution on the set of all possible CDF's is, to put it mildly, harder. Specifying distributions on function spaces is the task of Bayesian nonparametric modeling. In this paper, we will demonstrate one possible specification for the modeling of health care claims costs. As we will point out in the paper, prediction is especially problematic if there is misspecification of the prior distributions. Nonparametric methodology can be especially helpful if there is some unanticipated structure in the distribution of the parameters.

We first describe the data set that will be used. Next, we will specify the mathematical structure of the models in the full parametric and nonparametric settings. We provide more detail for the nonparametric setting since the parametric formulation is more familiar. Besides the mathematical detail, we also provide the algorithms necessary to implement the nonparametric model in an appendix. We present a small simulation study to further describe how the various models work. That is, we will demonstrate model performance when the truth is known. Finally, we present results from analyses of the 1994 data using the two model specifications, and rate them by evaluating their performance in predicting costs in 1995.

2 The Data

The data set is from a major medical plan, covering a block of medium sized groups in Illinois and Wisconsin for 1994 and 1995. Each policy holder was part of a group plan. In 1994 the groups consisted of from 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 employee plus one individual (often employee plus spouse) type. Table 1 gives some summary information about costs per day in 1994 and 1995.

Though the data are dated from a business perspective, they provide the ability to examine these two analysis paradigms without divulging proprietary information.

Insert Table 1 about here

Costs were assigned to each policyholder on a yearly basis and not assigned by episode of care or by medical incident. The costs were total costs, with deductible and co-payments added back in. The total yearly costs were then divided by the number of days of exposure. 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 number of days of exposure were small enough that total costs did not exceed $25,000.

Data consist of claims costs per day of exposure by policyholder. While age and gender of policyholder were known, age and gender of the claimant were only known when the claimant was the policyholder. We also did not know if multiple claims were made on the policy by the same individual or different individuals covered by the policy during the year. Knowledge and use of additional policy and claimant specific data would improve prediction. However, such information would also make the presentation more difficult to follow with the additional detail. We use the data available to present an expository illustration. The methods shown may be extended to more involved data sets.

3 The Models

3.1 The Hierarchical Parametric Bayes Model

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. Propensity, 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, representing no claims. We chose to deal with this by having a likelihood with a point mass at zero with probability π_i for group i. The parameter π_i depends on the group membership. Severity, the cost of a claim given that a claim is paid, is positively skewed. We chose 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) 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 feel that with the added richness available from the nonparametric model, the gamma likelihood should be sufficient 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,

$$f(y;\gamma,\theta) = \left(\frac{1}{(\theta)^{(\gamma)}\Gamma(\gamma)} y^{\gamma-1}\exp(\frac{-y}{\theta})\right) \tag{2}$$

The likelihood follows using a compound distribution argument:

$$\prod_{i=1}^{I}\prod_{\ell=1}^{L_i}\left[\pi_{i[y_{i\ell}=0]} + (1-\pi_i) f(y_{i\ell};\gamma_i,\theta_i)_{[y_{i\ell}>0]}\right], \tag{3}$$

where i indexes the group number, I 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 propensity parameter for group i, θ_i and γ_i are the severity parameters for group i, and $y_{i\ell}$ is the cost per day of exposure for each policyholder. Thus, we have a point mass probability for $y_{i\ell} = 0$, and a gamma likelihood for $y_{i\ell} > 0$.

The assignment of prior distributions should be a critical part of any analysis. One of the strengths of the full Bayesian approach is the ability the analyst has 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 case 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 propensity parameters π_i and the severity parameters (γ_i, θ_i). Conditionally on hyperparameters, we assume independent components. In particular,

$$\begin{aligned} \pi_i \mid \mu_\pi &\overset{\text{ind.}}{\sim} \text{Beta}(\mu_\pi, s_\pi^2), \quad i = 1, ..., I \\ \gamma_i \mid \beta &\overset{\text{ind.}}{\sim} \text{Gamma}(b, \beta), \quad i = 1, ..., I \\ \theta_i \mid \delta &\overset{\text{ind.}}{\sim} \text{Gamma}(d, \delta), \quad i = 1, ..., I. \end{aligned} \tag{4}$$

Here, to facilitate prior specification, we work with the Beta distribution parametrized in terms of its mean μ_π and variance s_π^2, i.e., with density given by

$$\frac{1}{\text{Be}(c_1, c_2)} \pi^{c_1 - 1} (1 - \pi)^{c_2 - 1}, \quad \pi \in (0, 1), \tag{5}$$

where $c_1 = s_\pi^{-2}(\mu_\pi^2 - \mu_\pi^3 - \mu_\pi s_\pi^2)$, $c_2 = s_\pi^{-2}(\mu_\pi - 2\mu_\pi^2 + 3\mu_\pi^3 - s_\pi^2 + \mu_\pi s_\pi^2)$, and $\text{Be}(\cdot, \cdot)$ denotes the Beta function, $\text{Be}(r, t) = \int_0^1 u^{r-1}(1 - u)^{t-1} \text{d}u$, $r > 0$, $t > 0$ (Evans, *et al.* 2000). We fix the hyperparameters s_π^2, b and d and assign reasonably non-informative priors to μ_π, β and δ. Specifically, 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 s_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, $(\{(\pi_i, \gamma_i, \theta_i) : i = 1, ..., I\}, \mu_\pi, \beta, \delta)$, is then proportional to

$$\begin{aligned} p(\mu_\pi) p(\beta) p(\delta) \left[\prod_{i=1}^{I} \frac{\beta^{-b}}{\Gamma(b)} \gamma_i^{b-1} \exp(\frac{-\gamma_i}{\beta}) \frac{\delta^{-d}}{\Gamma(d)} \theta_i^{d-1} \exp(\frac{-\theta_i}{\delta}) \frac{1}{\text{Be}(c_1, c_2)} \pi_i^{c_1 - 1} (1 - \pi_i)^{c_2 - 1} \right] \\ \left[\prod_{i=1}^{I} \prod_{\ell=1}^{L_i} \left\{ \pi_{i[y_{i\ell}=0]} + (1 - \pi_i) \left(f(y_{i\ell}; \gamma_i, \theta_i) \right)_{[y_{i\ell}>0]} \right\} \right], \end{aligned} \tag{6}$$

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

Current methods to analyze such a model include implementation of Markov chain Monte Carlo (MCMC) to produce samples from the posterior distributions which can then be evaluated (Gilks, *et al.* (1995)). MCMC is essentially Monte Carlo integration using Markov chains. Monte Carlo integration draws samples from the required distribution, and then forms sample averages to approximate expectations. MCMC draws these samples by running

a Markov chain for a long time. There are many ways of constructing these chains, but all of them are special cases of the general framework of Metropolis, *et al.* (1953) and Hastings (1970). Loosely speaking, the MCMC process draws samples from the posterior distributions by sampling throughout the appropriate support in the correct proportions. This is done using a Markov chain with the posterior as its stationary distribution.

More precisely, we first formulated the posterior distribution of each parameter, conditional on the other parameters, and assigned an initial value to each parameter. Then a new value was drawn from a "proposal" distribution. The ratio of the values of the complete conditionals computed using the proposed value and the old value of the parameters was computed and compared to a random uniform variate. If the ratio exceeded the random uniform, the proposed value was kept, otherwise the old value was kept. Using this method on each parameter, and cycling through the parameters, yielded a distribution that converged to the appropriate posterior for each parameter. For a more complete exposition of this methodology, the interested reader should refer to Scollnik (2001) or Gilks (1995). We then take the posterior draws for the paramaters to produce estimators such as means, quantiles, variances, etc.

To draw new parameters, we essentially deal with 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 $(\gamma_{\text{new}}, \theta_{\text{new}}, \pi_{\text{new}})$ from their respective prior distributions given the current values of the hyperparameters. Thus, draws of new parameters are dependent on the form of the prior distributions. The consequence is that if the prior distributions are misspecified, draws of new parameters will not mirror the actual setting. Estimating 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 for estimating costs is that those costs arising from groups that may be present in the future, but not being modeled with the current data, will not be accurate if the prior specification of the distribution of the parameters is not reflective of the 'truth'. We demonstrate the impact of this idea in Section 5.

3.2 The Nonparametric Bayes Model

The parametric random-effects distributions chosen for the π_i, γ_i and θ_i in Section 3.1 are modeling choices that might not be appropriate for specific data sets. Moreover, since these are distributions for latent model parameters, it is not straightforward to anticipate their form and/or shape based on exploratory data analysis. Bayesian nonparametric methods provide a flexible approach to handle this problem. The key idea is to use a *nonparametric* prior model for the random-effects distributions that supports essentially all possible distribution shapes, which at the same time can be *centered* around familiar parametric forms enabling relatively simple prior specification. Then, through the prior to posterior updating, the data are allowed to drive the shape of the posterior predictive estimates for the random-effects distributions. And this shape can be quite different from standard parametric forms (when these forms are not supported by the data), thus 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 achieves the goals discussed above. 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); Walker, *et al.*, (1999); Müller and Quintana, (2004); and Hanson, *et al.*, (2005).

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 $(\gamma_i, \theta_i, \pi_i)$, we believed the forces acting on the severity parameters could well have been different than those acting on the propensity parameters, so we have chosen to treat these parameters separately. Thus, we have a DP prior for the random-effects distribution, G_1, associated with the π_i as well as a separate (independent) DP prior for the random-effects distribution, G_2, corresponding to the (γ_i, θ_i).

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

$$\begin{aligned} y_{i\ell} \mid \pi_i, \gamma_i, \theta_i &\overset{\text{ind.}}{\sim} \pi_{i[y_{i\ell}=0]} + (1-\pi_i) f(y_{i\ell}; \gamma_i, \theta_i)_{[y_{i\ell}>0]}, \\ &\quad \ell = 1, ..., L_i; \;\; i = 1, ..., I \\ \pi_i \mid G_1 &\overset{\text{i.i.d.}}{\sim} G_1, \;\; i = 1, ..., I \\ (\gamma_i, \theta_i) \mid G_2 &\overset{\text{i.i.d.}}{\sim} G_2, \;\; i = 1, ..., I \\ G_1, G_2 &\overset{\text{ind.}}{\sim} \text{DP}(\alpha_1, G_{10}) \times \text{DP}(\alpha_2, G_{20}). \end{aligned} \tag{7}$$

Here, $\alpha_1, \alpha_2 > 0$ are the precision parameters of the DP priors, and G_{10} and G_{20} are the centering distributions. We set $G_{10}(\pi) = \text{Beta}(\pi; \mu_\pi, s_\pi^2)$, i.e., 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 s_π^2 to be fixed. For G_{20} we take independent Gamma components, i.e., $G_{20}((\gamma, \theta); \beta, \delta) = \text{Gamma}(\gamma; b, \beta) \times \text{Gamma}(\theta; d, \delta)$, 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 $\text{DP}(\alpha, G_0)$ prior, α controls how close a realization G is to G_0. In the DP mixture model in (7), the precision parameters control the distribution of the number of distinct elements I_1^* of the vector of $\{\pi_1, \ldots, \pi_I\}$ (controlled by α_1) and I_2^* of the vector $\{(\gamma_1, \theta_1), \ldots, (\gamma_I, \theta_I)\}$ (controlled by α_2), and hence, the number of distinct components of the mixtures. The number of distinct groups is, with positive probability, smaller than I, and, in fact, for typical choices of α_1 and α_2, fairly small relative to I. For instance, for moderate to large I,

$$E(I_k^* \mid \alpha_k) \approx \alpha_k \log\left(\frac{\alpha_k + I}{\alpha_k}\right), k = 1, 2, \tag{8}$$

and exact expressions for the prior probabilities $\Pr(I_k^* = m \mid \alpha_k)$, $m = 1, \ldots, I$, 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 (7).

3.2.1 Posterior Inference

To obtain posterior inference, we work with the marginalized version of model (7), which results by integrating G_1 and G_2 over their independent DP priors,

$$\begin{array}{rcl} y_{i\ell} \mid \pi_i, \gamma_i, \theta_i & \overset{\text{ind.}}{\sim} & \pi_{i[y_{i\ell}=0]} + (1-\pi_i) f(y_{i\ell}; \gamma_i, \theta_i)_{[y_{i\ell}>0]}, \\ & & \ell = 1, ..., L_i; \;\; i = 1, ..., I \\ (\pi_1, ..., \pi_I) \mid \mu_\pi & \sim & p(\pi_1, ..., \pi_I \mid \mu_\pi) \\ (\gamma_1, \theta_1), ..., (\gamma_I, \theta_I) \mid \beta, \delta & \sim & p((\gamma_1, \theta_1), ..., (\gamma_I, \theta_I) \mid \beta, \delta), \\ \beta, \delta, \mu_\pi & \sim & p(\beta) p(\delta) p(\mu_\pi), \end{array} \tag{9}$$

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

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

and $p((\gamma_1, \theta_1), ..., (\gamma_I, \theta_I) \mid \beta, \delta)$ is given by

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

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 $i = 1, ..., I$,

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

and

$$p((\gamma_i, \theta_i) \mid \{(\gamma_j, \theta_j) : j \neq i\}, \beta, \delta) = \frac{\alpha_2}{\alpha_2 + I - 1} g_{20}((\gamma_i, \theta_i); \beta, \delta) + \frac{1}{\alpha_2 + I - 1} \sum_{j=1}^{I-1} \delta_{(\gamma_j, \theta_j)}(\gamma_i, \theta_i). \tag{11}$$

Intuitively, the idea for posterior sampling using expressions (10) and (11), is that proposal values for the parameters are drawn from either the centering distribution or from values for previous draws of the other parameters ($j \neq i$). 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. For specific details concerning implementation of the MCMC algorithm in this nonparametric model, we refer the interested reader to Appendix B.

3.2.2 Posterior Predictive Inference

We will focus on the posterior predictive distribution for a new group. Denote by y_{new} the cost for a (new) policyholder within a new group. To obtain $p(y_{\text{new}} \mid \text{data})$, we need the posterior predictive distributions for a new π_{new} and for a new pair $(\gamma_{\text{new}}, \theta_{\text{new}})$. Let $\boldsymbol{\phi}$ be the full parameter vector corresponding to model (9), i.e., $\boldsymbol{\phi} = \{\pi_1, ..., \pi_I, (\gamma_1, \theta_1), ..., (\gamma_I, \theta_I), \beta, \delta, \mu_\pi\}$.

To obtain the expressions for $p(\pi_{\text{new}} \mid \text{data})$, $p((\gamma_{\text{new}}, \theta_{\text{new}}) \mid \text{data})$ and $p(y_{\text{new}} \mid \text{data})$, we need an expression for $p(y_{\text{new}}, \pi_{\text{new}}, (\gamma_{\text{new}}, \theta_{\text{new}}), \boldsymbol{\phi} \mid \text{data})$. This results by adding y_{new} to the first stage of model (7), and π_{new} and $(\gamma_{\text{new}}, \theta_{\text{new}})$ to the second and third stages of model (7), and then again marginalizing G_1 and G_2 over their DP priors. Specifically,

$$\begin{aligned} p(y_{\text{new}}, \pi_{\text{new}}, (\gamma_{\text{new}}, \theta_{\text{new}}), \boldsymbol{\phi} \mid \text{data}) &= \{\pi_{\text{new}[y_{\text{new}}=0]} + (1 - \pi_{\text{new}}) \\ &\quad \times f(y_{\text{new}}; \gamma_{\text{new}}, \theta_{\text{new}})_{[y_{\text{new}}>0]}\} \\ &\quad \times p((\gamma_{\text{new}}, \theta_{\text{new}}) \mid (\gamma_1, \theta_1), ..., (\gamma_I, \theta_I), \beta, \delta) \\ &\quad \times p(\pi_{\text{new}} \mid \pi_1, ..., \pi_I, \mu_\pi) \times p(\boldsymbol{\phi} \mid \text{data}), \end{aligned} \tag{12}$$

where

$$p(\pi_{\text{new}} \mid \pi_1, ..., \pi_I, \mu_\pi) = \frac{\alpha_1}{\alpha_1 + I} g_{10}(\pi_{\text{new}}; \mu_\pi, s_\pi^2) + \frac{1}{\alpha_1 + I} \sum_{i=1}^{I} \delta_{\pi_i}(\pi_{\text{new}}) \tag{13}$$

and

$$\begin{aligned} p((\gamma_{\text{new}}, \theta_{\text{new}}) \mid (\gamma_1, \theta_1), ..., (\gamma_I, \theta_I), \beta, \delta) = \frac{\alpha_2}{\alpha_2 + I} g_{20}((\gamma_{\text{new}}, \theta_{\text{new}}); \beta, \delta) + \\ \frac{1}{\alpha_2 + I} \sum_{i=1}^{I} \delta_{(\gamma_i, \theta_i)}(\gamma_{\text{new}}, \theta_{\text{new}}). \end{aligned} \tag{14}$$

Now, using the posterior samples for $\boldsymbol{\phi}$ (resulting from the MCMC algorithm described in Appendix B) and with appropriate integrations in expression (12), we can obtain posterior predictive inference for π_{new}, $(\gamma_{\text{new}}, \theta_{\text{new}})$, and y_{new}. In particular,

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

and therefore posterior predictive draws for π_{new} can be obtained by drawing from (13) for each posterior sample for $\pi_1, ..., \pi_I, \mu_\pi$. Moreover,

$$p((\gamma_{\text{new}}, \theta_{\text{new}}) \mid \text{data}) = \int p((\gamma_{\text{new}}, \theta_{\text{new}}) \mid (\gamma_1, \theta_1), ..., (\gamma_I, \theta_I), \beta, \delta) p(\boldsymbol{\phi} \mid \text{data}) \text{d}\boldsymbol{\phi},$$

can be sampled by drawing from (14) for each posterior sample for $(\gamma_1, \theta_1), ..., (\gamma_I, \theta_I), \beta, \delta$. Finally,

$$\begin{aligned} p(y_{\text{new}} \mid \text{data}) &= \int\int\int \{\pi_{\text{new}[y_{\text{new}}=0]} + (1 - \pi_{\text{new}}) f(y_{\text{new}}; \gamma_{\text{new}}, \theta_{\text{new}})_{[y_{\text{new}}>0]}\} \\ &\quad \times p(\pi_{\text{new}} \mid \pi_1, ..., \pi_I, \mu_\pi) \\ &\quad \times p((\gamma_{\text{new}}, \theta_{\text{new}}) \mid (\gamma_1, \theta_1), ..., (\gamma_I, \theta_I), \beta, \delta) \\ &\quad \times p(\boldsymbol{\phi} \mid \text{data}) \ \text{d}\pi_{\text{new}} \, \text{d}(\gamma_{\text{new}}, \theta_{\text{new}}) \, \text{d}\boldsymbol{\phi}. \end{aligned}$$

Based on this expression, posterior predictive samples for y_{new} can be obtained by first drawing π_{new} and $(\gamma_{\text{new}}, \theta_{\text{new}})$ (using expressions (13) and (14), respectively, for each posterior sample for $\boldsymbol{\phi}$) and then drawing y_{new} from $\pi_{\text{new}[y_{\text{new}}=0]} + (1 - \pi_{\text{new}}) f(y_{\text{new}}; \gamma_{\text{new}}, \theta_{\text{new}})_{[y_{\text{new}}>0]}$. Therefore, the posterior predictive distribution for a new group will have a point mass at 0 (driven by the posterior draws for π_{new}) and a continuous component (driven by the posterior draws for $(\gamma_{\text{new}}, \theta_{\text{new}})$).

Expressions (13) and (14) highlight the clustering structure induced by the DP priors, which enables flexible data-driven shapes in the posterior predictive densities $p(\pi_{\text{new}} \mid \text{data})$ and $p(\gamma_{\text{new}}, \theta_{\text{new}} \mid \text{data})$, and thus, flexible tail behavior for the continuous component of $p(y_{\text{new}} \mid \text{data})$. The utility of such flexibility in the prior is illustrated in the following sections with both the simulated and the real data.

4 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 drew parameters from unimodal distributions, and one drew 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 data were generated 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 nonparametric models.

In Case I (the unimodal case), the γ_i were drawn from a Gamma$(2, 5)$, the θ_i from a Gamma$(2, 10)$, and the π_i from a Beta$(4, 5)$. The draws were independent, and given these parameters, the data were drawn according to the likelihood in (3).

In Case II (the multimodal case), the γ_i were drawn from either a Gamma$(2, 1)$ or a Gamma$(50, 1)$ with equal probability. The θ_i were drawn independently using the same scenario as the γ_i, and the π_i 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 (3).

The parametric model was fit using the paradigm outlined previously. We ultimately chose $s_\pi^2 = 0.03$, $b = d = 1$, 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.

There are two main messages to be taken from the simulation results. One is that posterior point estimation of parameters for the groups represented in the simulated data sets is quite similar for the two models. In Figures 1, 2, and 3, we show posterior intervals (5^{th} to 95^{th} percentiles) for each group in simulation Case II. It is clear that both methods separate the modes in the prior densities quite well for the estimated parameters.

The second message is that the parametric model might not replicate the modes when predicting parameters for new groups, while the nonparametric methodology performs quite well at this task. Figures 4 and 5 demonstrate this. In Figure 4 we see the results from Case I, the unimodal case. The posterior predictive densities from the parametric model follow the generated parameter histograms quite closely. The nonparametric model produces comparable results. However, in Figure 5 it is obvious that the parametric model cannot predict the multiple modes. The nonparametric model does quite well at this task since the prior distributions are covered by the DP priors. This result 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), 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.

Referring back to Figures 1, 2, and 3, it is of interest that the posterior intervals are generally wider for the parametric model. This may also be explained by examining Figure 5. 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. It is ironic that artificially high certainty regarding the prior distributions of the parameters can lead to artificially high uncertainty regarding the parameter estimates.

5 Analysis of the Claims Data

The 1994 data consisted of $8,921$ observations in $1,075$ groups. Because of work with other policy types, we expected the γ_i to be smaller with the actual data than 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 10^{th}. Both models displayed convergent chains for the posterior draws of all parameter densities (Raftery and Lewis, 1995, and Smith, 2001).

In Figure 6 we show posterior predictive densities for both the parametric and nonparametric models for the γ_{new}, θ_{new}, and π_{new}. We note that the nonparametric model posterior predictive densities showed multimodal behavior like that we demonstrated in Case II of the simulation study. As in the simulation study, the parametric model cannot reveal this kind of behavior. When 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, a likely overestimation of uncertainty in the parameter estimates under the parametric model, as was displayed in the simulation. We also want to 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 by examining the 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 in 1995. We had no way to determine how many members were the same in both years. We obtained the posterior predictive distribution for this group using both models, using posterior samples from the corresponding triple $(\pi_i, \gamma_i, \theta_i)$. In Figure 7 (left panel), we show the posterior predictive distribution for the non-zero data for both the parametric and the nonparametric model as well as the histogram of the actual 1995 non-zero data for that group. There is little difference in the posterior predictive distributions, and both model the 1995 data reasonably well.

Thus, we now focus on predicting outcomes in 1995 for groups not present in the 1994 data. There were $8,732$ observations in 1995, and 522 of these 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, using the approaches discussed in Section 3.1 and 3.2.2, respectively. In Figure 7 (right panel), we show the posterior predictive densities for non-zero data from a new group using both the parametric and nonparametric models as well as a histogram of the actual 1995 data. We can see 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.

To further quantify the differences between the posterior predictive distributions, we computed a posterior predictive model comparison criterion. If $y_{0j}, j = 1, \ldots, J$, represents the positive observations from all new groups in 1995, we can estimate $p(y_{0j} \mid \text{data})$ (i.e., the conditional predictive ordinate (cpo)) at y_{0j} using

$$B^{-1} \sum_{b=1}^{B} f(y_{0j}; \gamma_{\text{new},b}, \theta_{\text{new},b}), \tag{15}$$

where $\{(\gamma_{\text{new},b}, \theta_{\text{new},b}) : b = 1, \ldots, B\}$ is the sample from the posterior predictive distribution for $(\gamma_{\text{new}}, \theta_{\text{new}})$ ($B = 10,000$ in our analysis). Of the $J = 371$ non-zero observations in 1995, 327 cpo's were greater for the nonparametric model. These values can also be summarized using the "cross-validation posterior predictive criterion", which is given by

$$Q(M_k) = J^{-1} \sum_{j=1}^{J} \log(p(y_{0j} \mid \text{data})) \tag{16}$$

where M_1 is the parametric model and M_2 is the nonparametric model. For the parametric model the value of Q was -3.20 while for the nonparametric model $Q = -2.94$. Thus, the predictive ability of the nonparametric model exceeded that of the parametric model for these data. Again, given the multimodal nature of the posterior predictive distributions for the $\pi_{\text{new}}, \gamma_{\text{new}}$, and θ_{new}, we are not surprised by this outcome.

6 Discussion

Bayesian nonparametric methods provide the practitioner with a class of models that offer real advantages when it comes to prediction. The idea of Bayesian nonparametrics is to place

prior distributions on spaces of functions, rather than 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 might be anticipated by the statistician.

In the data we examined, it is not unreasonable to believe that there might be multiple modes. 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, we would expect to see longer tails in these groups. Some small proportion of the groups might be extremely long in the tails. Looking at Figure 6 we can see 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 the means of the groups are not moving a great deal, while the variances for the groups with a few individuals with worse health is quite a bit larger. This is the kind of behavior we might expect from groups whose individuals are somewhat older, and thus more susceptible to larger health care expenditures. So it may be that the need for the nonparametric model in this case was a result of not being able to include age in the model. The problem, of course, is that failing to measure important covariates is a common and ongoing issue.

While this association may seem obvious in retrospect, it is not something that would necessarily be obvious before completing the nonparametric 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.

We believe that the Bayesian nonparametric model 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 would be expected to pay high dividends in the insurance industry.

Bibliography

1. Antoniak, C.E. (1974). "Mixtures of Dirichlet processes with applications to nonparametric problems." *The Annals of Statistics*, 2:1152-1174.

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

3. Blei, D.M., and Jordan, M.I. (2006). "Variational inference for Dirichlet process mixtures." *Bayesian Analysis*, 1:121–144.

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

5. Bush, C.A., and MacEachern, S.N. (1996). "A semiparametric Bayesian model for randomized block designs." *Biometrika*, 83:275-285.

6. de Finetti, B. (1937). "La prévision: ses lois logiques, ses sources subjectives." *Ann. Inst. H. Poincaré, 7, 1-68.*

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

8. Dey, D., Mueller, P., and Sinha D. (1998). *Practical Nonparametric and Semiparametric Bayesian Statistics.* (Lecture Notes in Statistics, Volume 133) Springer Verlag, New York.

9. Escobar, M.D., and West, M. (1995). "Bayesian density estimation and inference using mixtures." *Journal of the American Statistical Association*, 90:205-215.

10. Evans, M., Hastings, N. and Peacock, B. (2000) *Statistical Distributions.* 3rd ed. J. Wiley and Sons, Inc., New York.

11. Fellingham, G.W., Tolley, H.D., and Herzog, T.N. (2005). "Comparing credibility estimates of health insurance claims costs." *North American Actuarial Journal*, 9(1):1-12.

12. Ferguson, T.S. (1973). "A Bayesian analysis of some non-parametric problems." *The Annals of Statistics*, 1:209-230.

13. Ferguson, T.S. (1974). "Prior distributions on spaces of probability measures." *The Annals of Statistics*, 2:615-629.

14. Ferguson, T.S. (1983). "Bayesian density estimation by mixtures of normal distributions." In *Recent Advances in Statistics* (Rizvi MH, Rustagi JS, Siegmund D, eds.), 287–302, Academic Press, New York.

15. Freedman, D.A. (1963). "On the asymptotic behavior of Bayes estimates in the discrete case." *Annals of Mathematical Statistics*, 34:1386-1403.

16. Gelfand A.E., and Kottas, A. (2002). "A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models." *Journal of Computational and Graphical Statistics*, 11:289–305.

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

18. Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1998). *Bayesian Data Analysis.* Chapman & Hall, London.

19. Gilks, W.R. (1995). "Full conditional distributions." In *Markov Chain Monte Carlo in Practice* (Gilks, W.R., Richardson, S., and Spiegelhalter, D.J., eds.), 75–88. Chapman & Hall, London.

20. Gilks, W.R., Richardson, S., and Spiegelhalter, D.J., eds. (1995). *Markov Chain Monte Carlo in Practice.* Chapman & Hall, London.

21. Griffin, J.E., and Steel, M.F.J. (2006). "Order-based dependent Dirichlet processes." *Journal of the American Statistical Association*, 101: 179-194.

22. Hanson, T., Branscum, A., and Johnson, W. (2005). "Bayesian nonparametric modeling and data analysis: An introduction." In *Handbook of Statistics, Volume 25: Bayesian Thinking, Modeling and Computation* (Dey, D.K., and Rao, C.R., eds.), 245–278, Elsevier, Amsterdam.

23. Hastings, W.K. (1970). "Monte Carlo sampling methods using Markov chains and their applications." *Biometrika*, 57:97-109.

24. Herzog, T.N. (1999). *Introduction to Credibility Theory.* ACTEX, Winsted, CT.

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

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

27. Klugman, S. (1992). *Bayesian Statistics in Actuarial Science with Emphasis on Credibility.* Kluwer, Boston.

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

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

30. Lo, A.Y. (1984). "On a class of Bayesian nonparametric estimates: I. Density estimates." *The Annals of Statistics*, 12:351-357.

31. MacEachern, S.N., and Müller, P. (1998). "Estimating mixtures of Dirichlet process models." *Journal of Computational and Graphical Statistics*, 7:223-238.

32. MacEachern, S.N., Clyde, M., and Liu, J.S. (1999). "Sequential importance sampling for nonparametric Bayes models: The next generation." *Canandian Journal of Statistics*, 27:251-267.

33. Makov, U. E. (2001). "Principal applications of Bayesian methods in actuarial science." *North American Actuarial Journal*, 5(4):53-73.

34. Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953). "Equation of state calculations by fast computing machine." *Journal of Chemical Physics*, 21:1087-1091.

35. Müller, P., and Quintana, F.A. (2004). "Nonparametric Bayesian data analysis." *Statistical Science*, 19:95-110.

36. Neal, R.M. (2000). "Markov chain sampling methods for Dirichlet mrocess mixture models." *Journal of Computational and Graphical Statistics*, 9:249-265.

37. Raftery, A.E., and Lewis, S.M. (1995). "Implementing MCMC." In *Markov Chain Monte Carlo in Practice* (Gilks, W.R., Richardson, S., and Spiegelhalter, D.J. eds.), 115-130. Chapman & Hall, London.

38. Sethuraman, J. (1994). "A constructive definition of Dirichlet priors." *Statistica Sinica*, 4:639-650.

39. Scollnik, D. (2001). "Actuarial modeling with MCMC and BUGS." *North American Actuarial Journal*, 5(2):96-124.

40. Smith, B.J. (2001). *Bayesian Output Analysis Program (BOA) Version 1.0.0 User's Manual.* http://www.public-health.uiowa.edu/boa/.

41. Venter, G.G. (1996). "Credibility." In *Foundations of Casualty Actuarial Science.* 375–483, Casualty Actuarial Society, Arlington, VA.

42. Walker, S., Damien, P., Laud, P., and Smith, A.F.M. (1999). "Bayesian nonparametric inference for random distributions and related functions (with discussion)." *Journal of the Royal Statistical Society Series B*, 61:485-527.

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); see also, e.g., Ferguson (1983), Lo (1984), Kuo (1986), and Brunner and Lo (1989) for early work on modeling and inference using DP mixtures.

The Dirichlet process. The DP (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 $\alpha[1 - G_0(x)]$ and, 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), i.e., a more flexible version of the DP prior that involves hyperpriors for α and/or the parameters $\boldsymbol{\psi}$ of $G_0(\cdot) \equiv G_0(\cdot|\boldsymbol{\psi})$.

A practically 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. $\mathrm{Beta}(1, \alpha)$; 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 $\mathrm{Beta}(1, \alpha)$ 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).

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

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

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 $\boldsymbol{Y} = (Y_1, ..., Y_n)$, e.g., 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:

$$\begin{array}{rcl} Y_i|\boldsymbol{\theta}_i & \overset{\text{ind.}}{\sim} & K(\cdot|\boldsymbol{\theta}_i), \;\; i = 1, ..., n \\ \boldsymbol{\theta}_i|G & \overset{\text{i.i.d.}}{\sim} & G, \;\; i = 1, ..., n \\ G \mid \alpha, \boldsymbol{\psi} & \sim & \mathrm{DP}(\alpha, G_0(\cdot|\boldsymbol{\psi})). \end{array} \tag{17}$$

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, e.g., $\boldsymbol{\phi}$ 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 (17) 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, e.g., the reviews in Müller and Quintana, 2004, and Hanson *et al.*, 2005).

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

Appendix B. – The MCMC Algorithm for the Nonparametric Model

The joint posterior, $p(\pi_1, ..., \pi_I, (\gamma_1, \theta_1), ..., (\gamma_I, \theta_I), \beta, \delta, \mu_\pi \mid \text{data})$, corresponding to model (9) is proportional to

$$p(\beta)p(\delta)p(\mu_\pi)p(\pi_1, ..., \pi_I \mid \mu_\pi)p((\gamma_1, \theta_1), ..., (\gamma_I, \theta_I) \mid \beta, \delta) \left\{ \prod_{i=1}^{I} \pi_i^{L_{i0}} (1 - \pi_i)^{L_i - L_{i0}} \right\} \left\{ \prod_{i=1}^{I} \prod_{\{\ell: y_{i\ell} > 0\}} f(y_{i\ell}; \gamma_i, \theta_i) \right\}, \tag{18}$$

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 (10) and (11) 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 $i = 1, ..., I$, 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 (10). 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 10)
- 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 $i = 1, ..., I$, the posterior full conditional for (γ_i, θ_i),

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

where $p((\gamma_i, \theta_i) \mid \{(\gamma_j, \theta_j) : j \neq i\}, \beta, \delta)$ is given by expression (11). 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 11)
- 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, $i = 1, ..., I$, are updated we obtain I_1^* ($\leq I$), the number of distinct π_i, and the distinct values π_j^*, $j = 1, ..., I_1^*$. Similarly, after updating all the (γ_i, θ_i), $i = 1, ..., I$, we obtain a number I_2^* ($\leq I$) of distinct (γ_i, θ_i) with distinct values (γ_j^*, θ_j^*), $j = 1, ..., I_2^*$.

Now, the posterior full conditional for β can be expressed as

$$p(\beta \mid ..., \text{data}) \propto \beta^{-3} \exp(-A_\beta/\beta) \times \prod_{j=1}^{I_2^*} \text{Gamma}(\gamma_j^*; b, \beta).$$

so

$$p(\beta \mid ..., \text{data}) \propto \beta^{-3} \exp(-A_\beta/\beta) \times \prod_{j=1}^{I_2^*} \beta^{-b} \exp(-\gamma_j^*/\beta) \propto \beta^{-(bI_2^*+3)} \exp(-(A_\beta + \sum_{j=1}^{I_2^*} \gamma_j^*)/\beta)$$

and we therefore recognize the posterior full conditional for β as an inverse gamma distribution with shape parameter $bI_2^* + 2$ and scale parameter $A_\beta + \sum_{j=1}^{I_2^*} \gamma_j^*$.

Analogously, the posterior full conditional for δ,

$$p(\delta \mid ..., \text{data}) \propto \delta^{-3} \exp(-A_\delta/\delta) \times \prod_{j=1}^{I_2^*} \text{gamma}(\theta_j^*; d, \delta),$$

and we therefore obtain an inverse gamma posterior full conditional distribution for δ with shape parameter $dI_2^* + 2$ and scale parameter $A_\delta + \sum_{j=1}^{I_2^*} \theta_j^*$.

Finally, the posterior full conditional for μ_π is given by

$$p(\mu_\pi \mid ..., \text{data}) \propto p(\mu_\pi) \times \prod_{j=1}^{I_1^*} g_{10}(\pi_j^*; \mu_\pi, s_\pi^2)$$

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

Table 1: Summary information for costs per day in dollars for 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

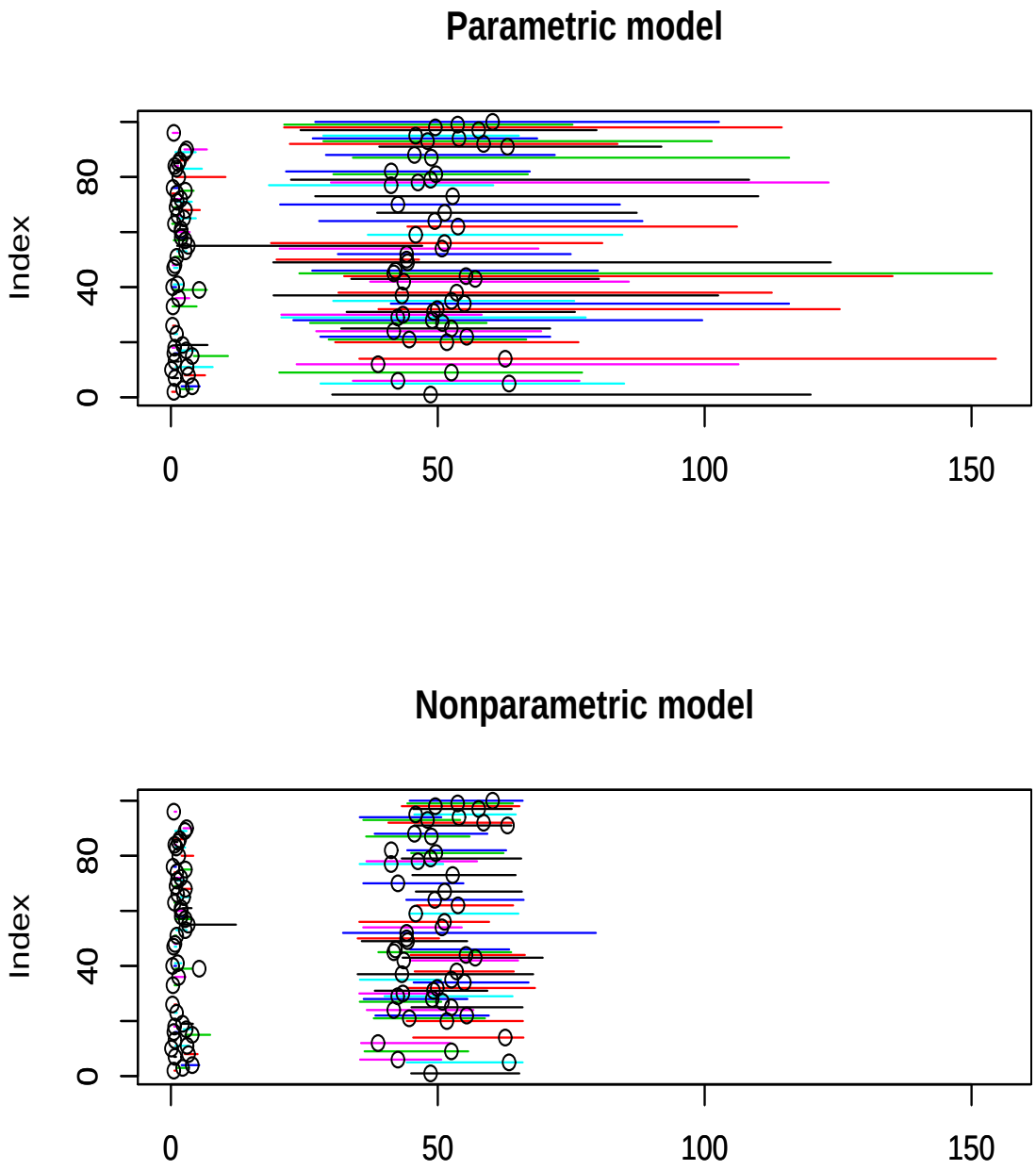

Figure 1: Simulation case II. Posterior intervals (5-th to 95-th posterior percentile) for each γ_i, $i = 1, ..., 100$, under the parametric (upper panel) and nonparametric (lower panel) models. The circles denote the actual generated γ_i.

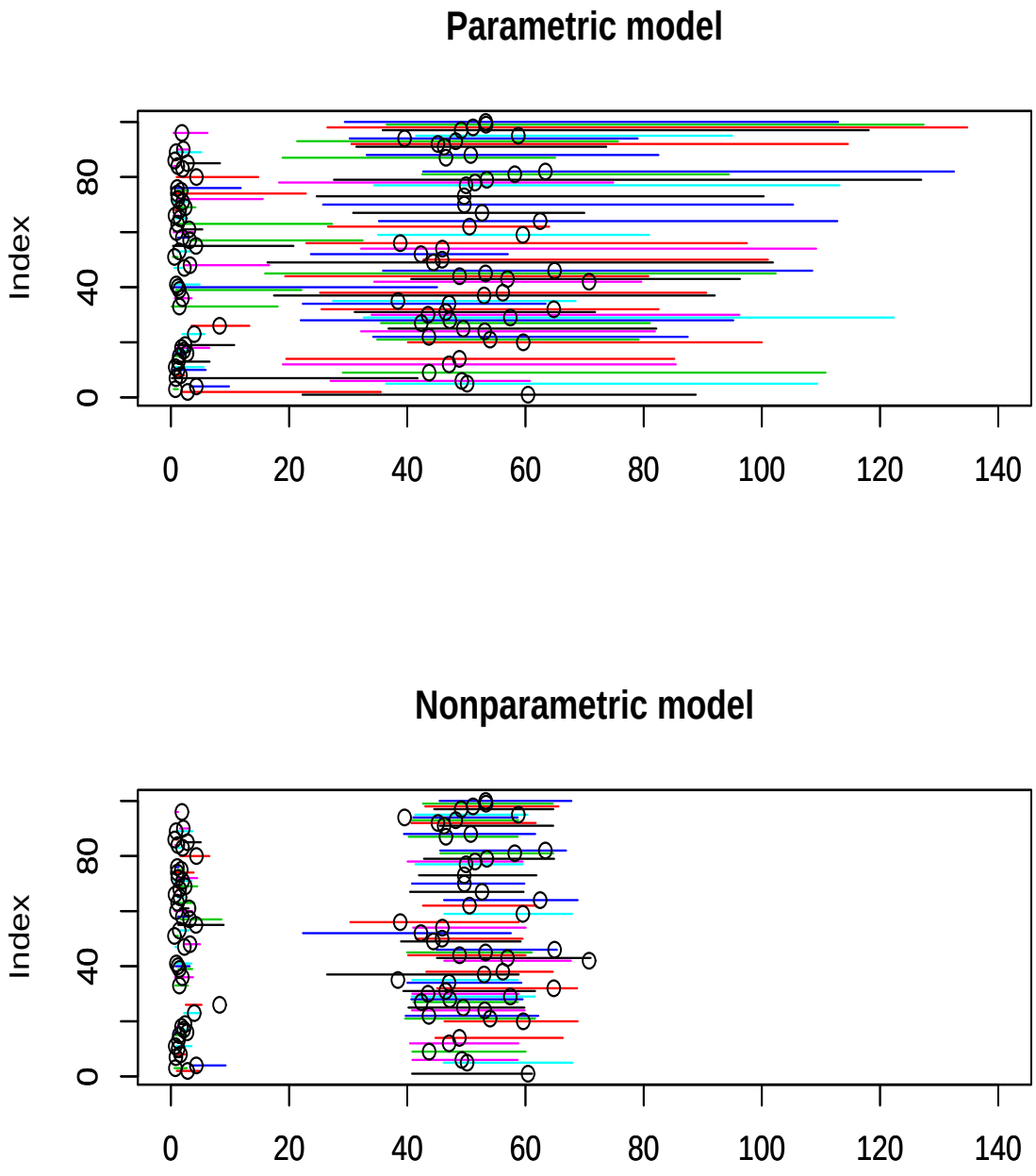

Figure 2: Simulation case II. Posterior intervals (5-th to 95-th posterior percentile) for each θ_i, $i = 1, ..., 100$, under the parametric (upper panel) and nonparametric (lower panel) models. The circles denote the actual generated θ_i.

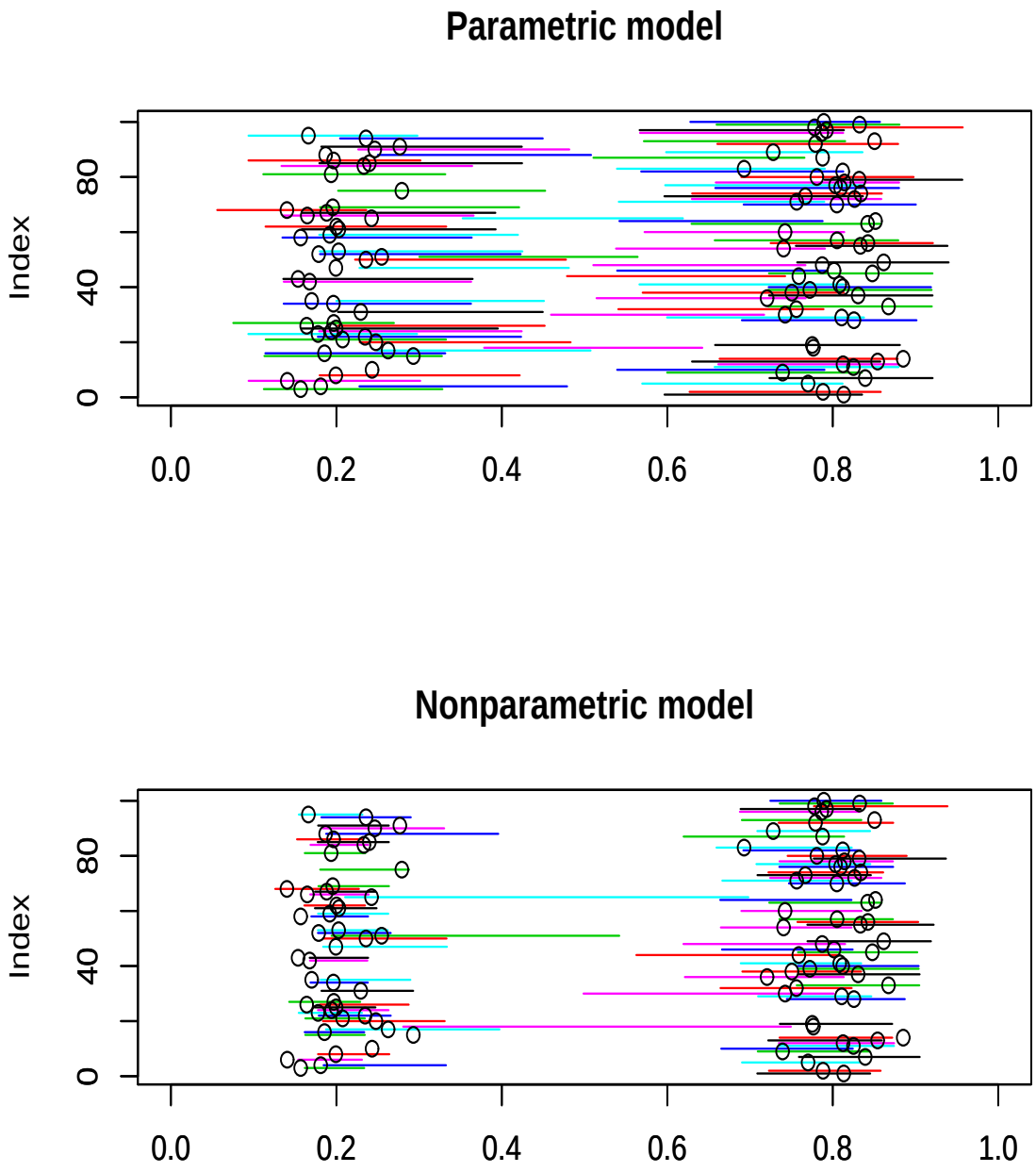

Figure 3: Simulation case II. Posterior intervals (5-th to 95-th posterior percentile) for each π_i, $i = 1, ..., 100$, under the parametric (upper panel) and nonparametric (lower panel) models. The circles denote the actual generated π_i.

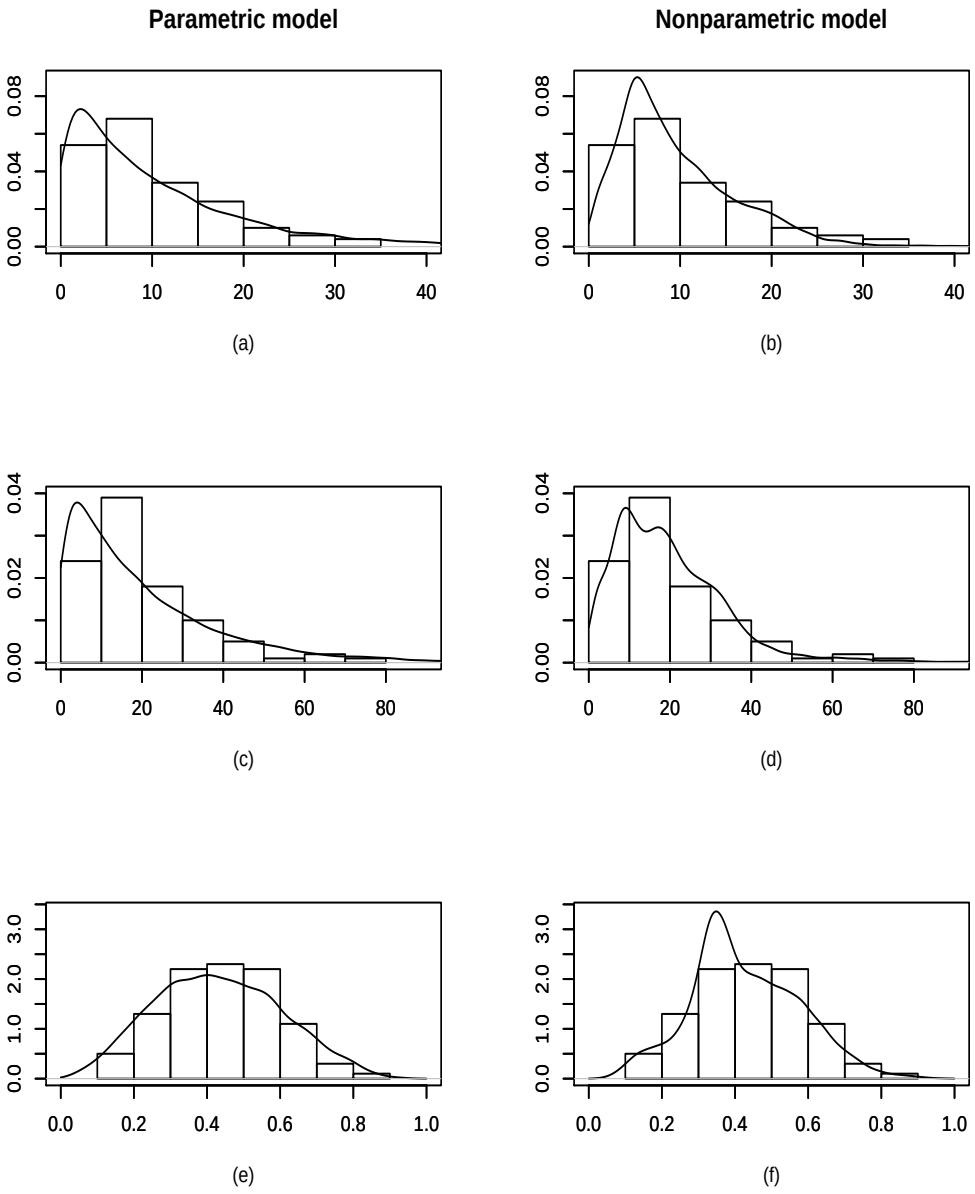

Figure 4: Simulation case I — unimodal data. Posterior predictive densities for γ_{new} (panels (a) and (b)), for θ_{new} (panels (c) and (d)), and for π_{new} (panels (e) and (f)), under the parametric model (left column) and the nonparametric model (right column). The histograms plot the generated γ_i (panels (a) and (b)), θ_i (panels (c) and (d)), and π_i (panels (e) and (f)), $i = 1, ..., 100$.

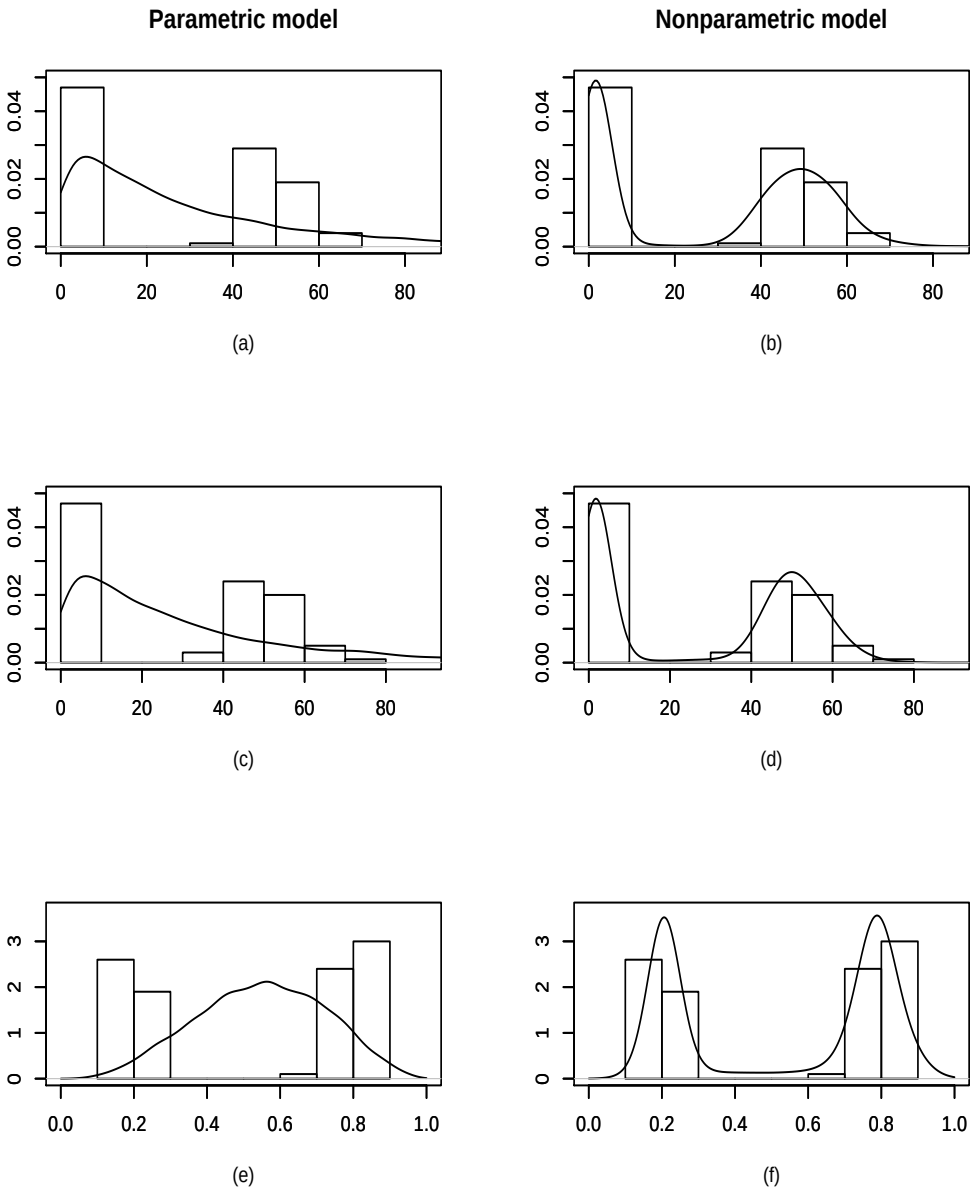

Figure 5: Simulation case II — multimodal data. Posterior predictive densities for γ_{new} (panels (a) and (b)), for θ_{new} (panels (c) and (d)), and for π_{new} (panels (e) and (f)), under the parametric model (left column) and the nonparametric model (right column). The histograms plot the generated γ_i (panels (a) and (b)), θ_i (panels (c) and (d)), and π_i (panels (e) and (f)), $i = 1, ..., 100$.

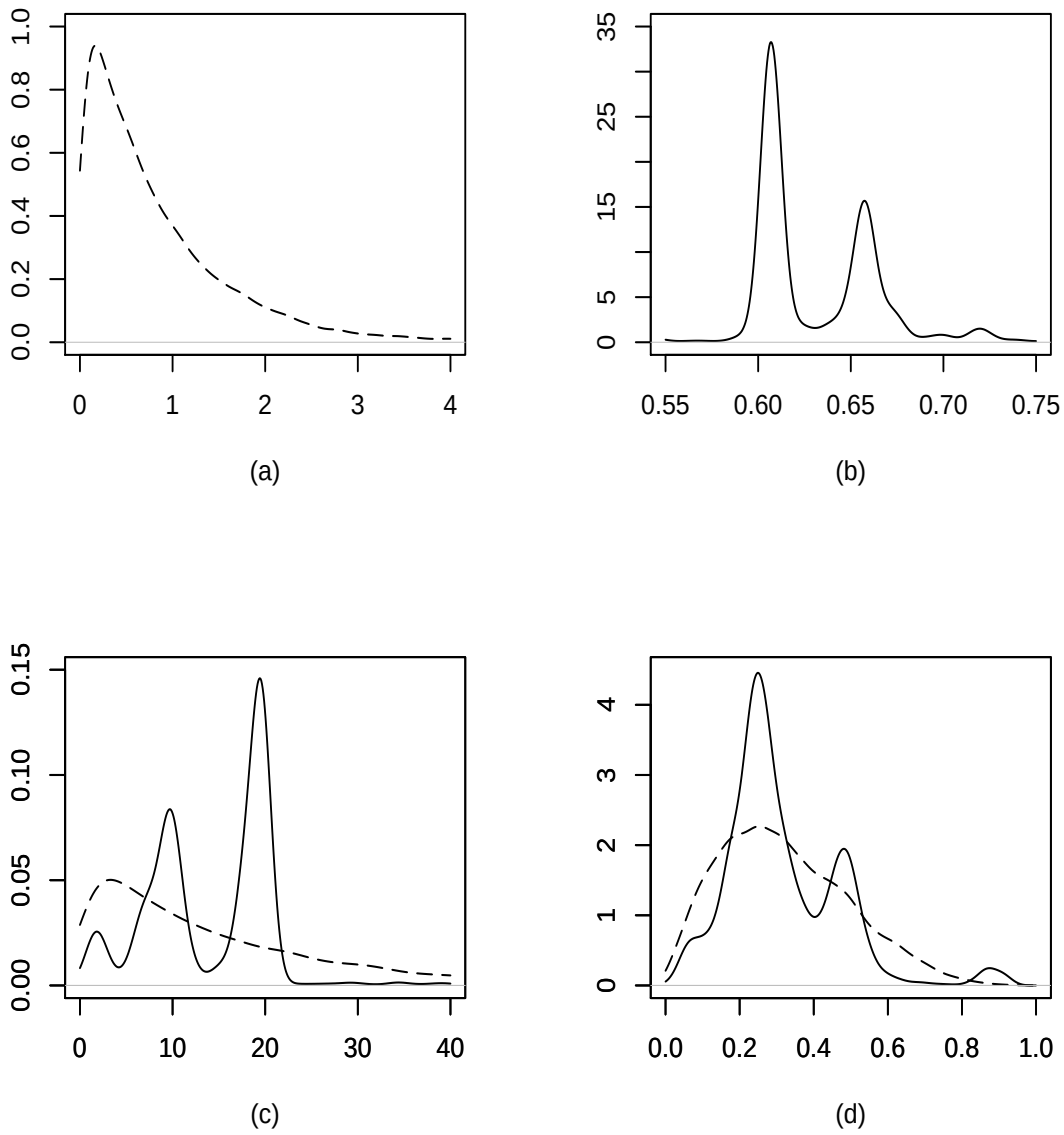

Figure 6: Posterior predictive inference for the random-effects distributions for the real data. Panels (a) and (b) include the posterior predictive density for γ_{new} under the parametric and nonparametric models, respectively. (Note the different scale in these two panels.) The posterior predictive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

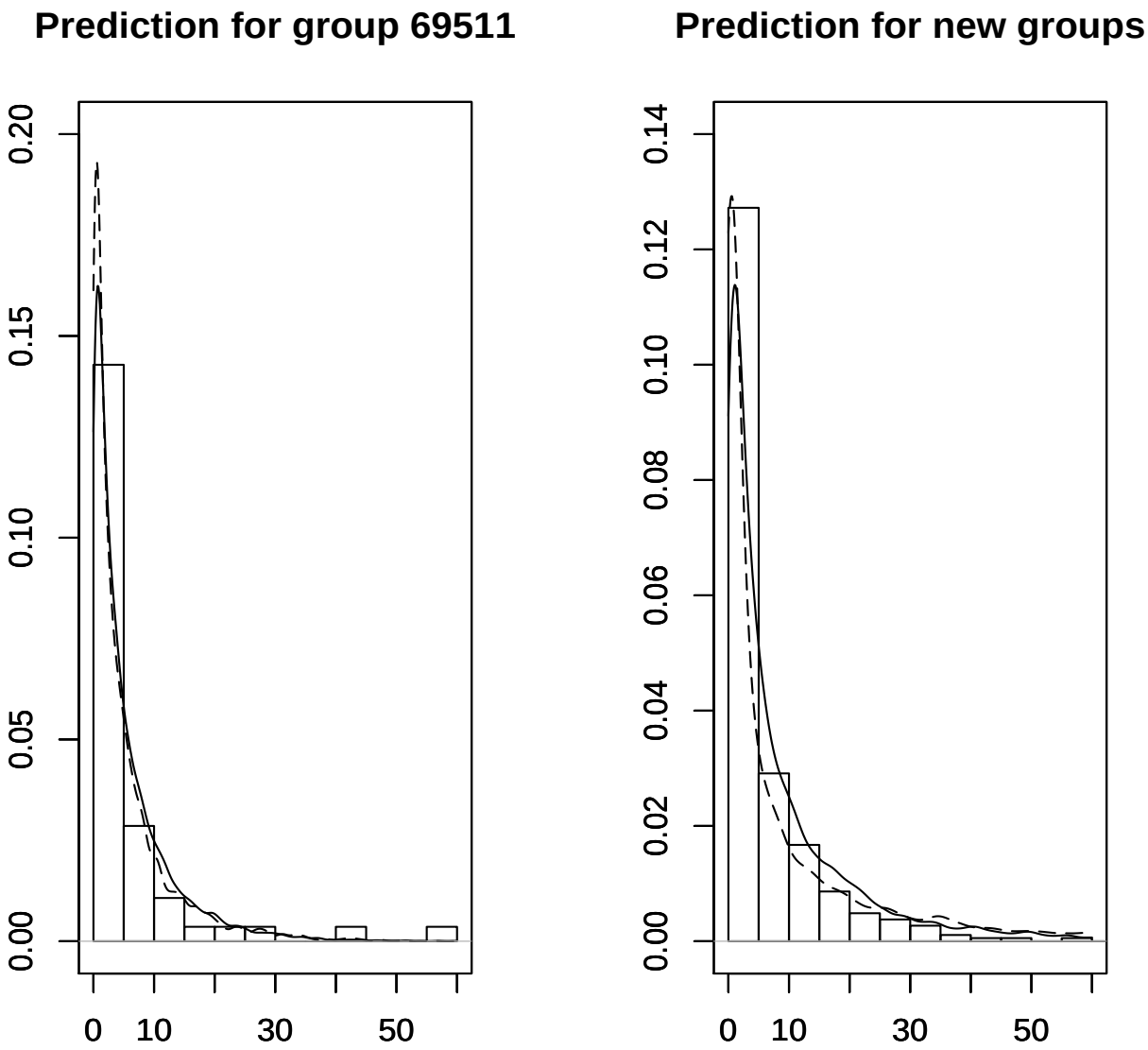

Figure 7: Cross-validated posterior predictive inference for the real data. Posterior results are based on data from year 1994 and are validated using corresponding data from year 1995 (given by the histograms in the two panels). The left panel includes posterior predictive densities for claims under 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.