Parametric and nonparametric Bayesian model specification: a case study involving models for count data

Milovan Krnjajić, Athanasios Kottas and David Draper
Department of Applied Mathematics and Statistics, Baskin School of Engineering, University of California, 1156 High Street, MS: SOE2, Santa Cruz, CA 95064, USA

Abstract In this paper we present the results of a simulation study to explore the ability of Bayesian parametric and nonparametric models to provide an adequate fit to count data, of the type that would routinely be analyzed parametrically either through fixed-effects or random-effects Poisson models. The context of the study is a randomized controlled trial with two groups (treatment and control). Our nonparametric approach utilizes several modeling formulations based on Dirichlet process priors. We find that the nonparametric models are able to flexibly adapt to the data, to offer rich posterior inference, and to provide, in a variety of settings, more accurate predictive inference than parametric models.

Keywords: Dirichlet process mixture model, Markov chain Monte Carlo methods, random-effects Poisson model, stochastically ordered distributions

1 Introduction

In an experiment conducted in the 1980s (Hendriksen et al. 1984), 572 elderly people living in a number of villages in Denmark were randomized, 287 to a control group, who received standard health care, and 285 to a treatment group, who received standard care plus *in-home geriatric assessment* (IHGA), a kind of preventive medicine in which each person's medical and social needs were assessed and acted upon individually. One important outcome was the number of hospitalizations during the two-year life of the study.

Table 1 presents the data. Because of the randomized controlled character of the study, it is reasonable to draw the causal conclusion that IHGA lowered the mean hospitalization rate per two years (for these elderly Danish people, at least) by $(0.944 - 0.768) \doteq 0.176$, which is about a 19% reduction from the control level, a difference that is large in clinical terms (and indeed Hendriksen et al. used this result to recommend widespread implementation of IHGA). So far this is simply description, combined with a judgment of practical

Table 1: *Distribution of number of hospitalizations in the IHGA study over a two-year period.*

	Number of Hospitalizations										
Group	0	1	2	3	4	5	6	7	n	Mean	Variance
Control	138	77	46	12	8	4	0	2	287	0.944	1.54
Treatment	147	83	37	13	3	1	1	0	285	0.768	1.02

significance. But what is the posterior distribution for the treatment effect in the entire population of patients judged exchangeable with those in the study? This is an inferential question, for which a statistical model is needed.

Since the outcome consists of counts of relatively rare events, Poisson modeling comes initially to mind; in the absence of strong prior information about the underlying hospitalization rates in the control and treatment groups, the first choice might well be a fixed-effects model of the form

$$C_i|\lambda^C \overset{\text{IID}}{\sim} \text{Poisson}(\lambda^C), \quad T_j|\lambda^T \overset{\text{IID}}{\sim} \text{Poisson}(\lambda^T),$$

for $i = 1, \ldots, n_C = 287$ and $j = 1, \ldots, n_T = 285$, with, say, a diffuse prior for (λ^C, λ^T). But the last two columns of Table 1 reveal that the sample variance-to-mean ratios in the control and treatment groups are 1.63 and 1.33, respectively, indicating substantial Poisson over-dispersion. The second parametric modeling choice might well therefore be a random-effects Poisson model of the form

$$\begin{array}{rcl} C_i|\lambda_i^C & \overset{\text{indep}}{\sim} & \text{Poisson}(\lambda_i^C), \;\; i = 1, ..., n_C \\ \log(\lambda_i^C) \mid \beta_0^C, \sigma_C^2 & \overset{\text{IID}}{\sim} & N(\beta_0^C, \sigma_C^2), \;\; i = 1, ..., n_C, \end{array} \tag{1}$$

and similarly for the treatment group. (Again, diffuse priors could be used for (β_0^C, σ_C^2).) This model is more scientifically satisfying: each patient in the control group has his/her own (latent) underlying rate of hospitalization λ_i^C, which may well differ from the underlying rates of the other control patients because of unmeasured differences in factors such as health status at the beginning of the experiment.

Model (1), when extended in parallel to the treatment group, specifies Lognormal mixtures of Poisson distributions as the implied sampling distributions of the hospitalization counts C_i and T_j, and is easy to fit via MCMC; the inferential question posed above is addressed in a straightforward way by monitoring the multiplicative effect parameter $\exp(\beta_0^T - \beta_0^C)$. However,

(a) nothing guarantees that the Gaussian mixing distribution in the last line of (1) is "correct," and moreover

(b) this model was arrived at by the usual data-analytic procedure in which we (i) enlist the aid of the data to specify the structural form of the model and then (ii) pretend that we knew all along that model (1) was appropriate.

As has been noted elsewhere by many observers (e.g., Draper 1995), this approach to model-building is both incoherent and liable to mis-calibration: we are in effect using the data twice, once to specify a prior distribution on structure space and then again to update this data-determined prior, and the result is likely to be inappropriately narrow uncertainty bands. Bayesian nonparametric (BNP) modeling, in which the mixing distribution is regarded as unknown—instead of dogmatically asserting that we somehow know it is Gaussian—may well provide a more satisfying approach to modeling data of this type. (See, e.g., Walker et al. 1999; Müller and Quintana 2004; Hanson, Branscum and Johnson 2005, for reviews of BNP modeling.) In this paper we contrast parametric and nonparametric models, based on Dirichlet process (DP) priors (Ferguson 1973), for over-dispersed count data, with the goal of exploring the practical consequences of nonparametric modeling as an alternative to the potentially mis-calibrated data-analytic approach to model-building.

We argue that such comparisons for generic statistical model settings are practically important to enhance our understanding of the performance of BNP models. The importance of this type of work has been recognized in the Bayesian literature, although a limited number of general studies appears to exist; see, e.g., Paddock et al. (2006) for a simulation study involving certain classes of hierarchical models with Gaussian first stage distributions, and the review papers by Müller and Quintana (2004) and Hanson, Branscum and Johnson (2005) for some related references to work that involves comparisons of parametric and nonparametric (or semiparametric) Bayesian models, typically, in the context of the analysis of specific data sets.

The plan of the paper is as follows. In Section 2 we specify the fixed- and random-effects parametric models we study, and in Section 3 we describe two BNP models that involve DP priors for the random-effects distributions. Section 4 presents the design and analysis of a simulation experiment to compare parametric and nonparametric modeling in the context of data structures like those in the IHGA case study (randomized controlled trials with treatment and control groups and a discrete response). In Section 5 we describe another BNP approach that involves modeling with DP priors directly on the scale of the count data. In Section 6 we discuss results obtained by fitting four models to the IHGA

data, and Section 7 concludes with a discussion. In an Appendix we give a brief account of the DP and its use in mixture modeling, and we provide details of the computational approaches to obtaining posterior distributions of various quantities of interest in our context.

2 Parametric models

To fix notation, let Y_{1i} be the integer-valued outcomes in the control group, with Y_{2i} as the corresponding values in the treatment group, and denote by Poisson(θ) the Poisson distribution (the cumulative distribution function (CDF) or the probability mass function, depending on the context) with mean $\lambda = \exp(\theta)$, $\theta \in R$.

As noted in Section 1, simple parametric modeling formulations for data sets like those in Table 1 include fixed-effects Poisson models, i.e., for $r = 1, 2$,

$$Y_{ri}|\theta_r \overset{\text{IID}}{\sim} \text{Poisson}(\theta_r), \;\; i = 1, ..., n_r \tag{2}$$

with independent priors G_{r0} for θ_r; and random-effects Poisson models, i.e., for $r = 1, 2$,

$$\begin{array}{ccc} Y_{ri}|\theta_{ri} & \overset{\text{indep}}{\sim} & \text{Poisson}(\theta_{ri}), \;\; i = 1, ..., n_r \\ \theta_{ri} & \overset{\text{IID}}{\sim} & G_{r0}, \;\; i = 1, ..., n_r \end{array} \tag{3}$$

again, with independent random effects distributions G_{r0}. Under both (2) and (3), a standard choice for G_{r0} would be a Normal distribution, N(μ_r, σ_r^2), $r = 1, 2$. Both models are typically completed by adding Normal and Inverse-Gamma hyperpriors on μ_r and σ_r^2, respectively. We shall refer to (2) and (3) as models $\mathcal{M}_0$ and $\mathcal{M}_1$, respectively.

Posterior predictive inference under model (3) is straightforward. Furthermore, posteriors of the random effects distributions $G_{r0} = \text{N}(\mu_r, \sigma_r^2)$, $r = 1, 2$, are directly determined by the posterior samples of μ_r and σ_r^2. Hence, inference for various functionals is readily available; for instance, the posterior of the mean functional, $E(Y|G_{r0}) = \int \exp(\theta) \text{dN}(\theta; \mu_r, \sigma_r^2) = \exp(\mu_r + 0.5\sigma_r^2)$, can be used to address one of the inferential questions of interest in the control-treatment setting (Section 6 provides an illustration with the IHGA data).

3 Bayesian nonparametric models

Here we consider two BNP extensions to the Poisson random effects model. We treat the random effects distributions as unknown, and use the DP as a prior probability model on the space of such distributions.

3.1 Independent priors for the random-effects distributions

The first nonparametric extension of parametric model (3) emerges by relaxing the normality (or any other parametric distributional) assumption for the random-effects distributions and instead placing DP priors on the associated spaces of CDFs. (See the Appendix for a brief review of the DP.) For $r = 1, 2$, we obtain the following DP mixture model:

$$\begin{array}{rcl} Y_{ri}|\theta_{ri} & \overset{\text{indep}}{\sim} & \text{Poisson}(\theta_{ri}),\ \ i = 1, ..., n_r \\ \theta_{ri}|G_r & \overset{\text{IID}}{\sim} & G_r,\ \ i = 1, ..., n_r \\ G_r|(\alpha_r, \mu_r, \sigma_r^2) & \sim & \text{DP}[\alpha_r G_{r0}(\mu_r, \sigma_r^2)] \\ (\alpha_r, \mu_r, \sigma_r^2) & \sim & p(\alpha_r)p(\mu_r)p(\sigma_r^2), \end{array} \tag{4}$$

where the DP priors for G_r are independent, the base distributions $G_{r0} = \text{N}(\mu_r, \sigma_r^2)$, and $p(\alpha_r)$, $p(\mu_r)$, and $p(\sigma_r^2)$ are the hyperpriors for the DP parameters. We shall refer to (4) as model $\mathcal{M}_2$. The Pólya urn characterization of the DP (Blackwell and MacQueen 1973) yields a useful *marginalized* version of (4) by integrating out G_r over its DP prior. Specifically, for $r = 1, 2$,

$$\begin{array}{rcl} Y_{ri}|\theta_{ri} & \overset{\text{indep}}{\sim} & \text{Poisson}(\theta_{ri}),\ \ i = 1, ..., n_r \\ (\theta_{r1}, \ldots, \theta_{rn_r})|(\alpha_r, \mu_r, \sigma_r^2) & \sim & p(\theta_{r1}, \ldots, \theta_{rn_r}|\alpha_r, \mu_r, \sigma_r^2) \\ (\alpha_r, \mu_r, \sigma_r^2) & \sim & p(\alpha_r)p(\mu_r)p(\sigma_r^2), \end{array}$$

where the prior for the random effects, $p(\theta_{r1}, \ldots, \theta_{rn_r}|\alpha_r, \mu_r, \sigma_r^2)$, is given by

$$g_{r0}(\theta_{r1}|\mu_r, \sigma_r^2) \prod_{i=2}^{n_r} \left\{ \frac{\alpha_r}{\alpha_r + i - 1} g_{r0}(\theta_{ri}|\mu_r, \sigma_r^2) + \frac{1}{\alpha_r + i - 1} \sum_{\ell=1}^{i-1} \delta_{\theta_{r\ell}}(\theta_{ri}) \right\},$$

with g_{r0} denoting the density of G_{r0}. This expression specifies the joint prior probability model for the latent $\theta_{r1}, \ldots, \theta_{rn_r}$ induced by the DP prior, and indicates that both parametric models (2) and (3) are limiting cases of the DP mixture model, arising when (for $r = 1, 2$) $\alpha_r \to 0^+$ and $\alpha_r \to \infty$, respectively. The nonparametric DP mixture model adds flexibility with regard to posterior predictive inference, since it allows data-driven clustering in the θ_{ri}.

This clustering in the prior of the θ_{ri} results from the discreteness of the random distribution G_r under the DP prior (Ferguson 1973; Blackwell 1973). For $r = 1, 2$, let n_r^* be the number of clusters (distinct elements) in the vector $(\theta_{r1}, \ldots, \theta_{rn_r})$ and denote by $\boldsymbol{\theta}_r^* = (\theta_{r\ell}^* : \ell = 1, \ldots, n_r^*)$ the vector of the distinct θ_{ri}. The vector of configuration indicators $\boldsymbol{s}_r = (s_{r1}, \ldots, s_{rn_r})$, defined by $s_{ri} = \ell$ if and only if $\theta_{ri} = \theta_{r\ell}^*$, $i = 1, \ldots, n_r$, determines the clusters. Let $n_{r\ell}$ be the size of cluster ℓ, i.e., $n_{r\ell} = |\{i : s_{ri} = \ell\}|, \ell = 1, \ldots, n_r^*$. We used Markov chain Monte Carlo (MCMC) algorithm 6

from Neal (2000) to obtain posterior samples from $p((\theta_{r1}, \ldots, \theta_{rn_r}), \alpha_r, \mu_r, \sigma_r^2 | \boldsymbol{Y}_r)$, equivalently, from $p(n_r^*, \boldsymbol{s}_r, \boldsymbol{\theta}_r^*, \alpha_r, \mu_r, \sigma_r^2 | \boldsymbol{Y}_r)$, where $\boldsymbol{Y}_r = (Y_{r1}, \ldots, Y_{rn_r})$.

The posterior predictive distribution for a future observable Y_r^{new} under the control ($r = 1$) or treatment ($r = 2$) conditions is given by

$$p(Y_r^{\text{new}} | \boldsymbol{Y}_r) = \iint \text{Poisson}(Y_r^{\text{new}} | \theta_r^{\text{new}}) p(\theta_r^{\text{new}} | \boldsymbol{\eta}_r) p(\boldsymbol{\eta}_r | \boldsymbol{Y}_r), \tag{5}$$

where, based on the Pólya urn structure for the DP,

$$p(\theta_r^{\text{new}} | \boldsymbol{\eta}_r) = \frac{\alpha_r}{\alpha_r + n_r} g_{r0}(\theta_r^{\text{new}} | \mu_r, \sigma_r^2) + \frac{1}{\alpha_r + n_r} \sum_{\ell=1}^{n_r^*} n_{r\ell}\, \delta_{\theta_{r\ell}^*}(\theta_r^{\text{new}}), \tag{6}$$

with $\boldsymbol{\eta}_r$ collecting the variables $(n_r^*, \boldsymbol{s}_r, \boldsymbol{\theta}_r^*, \alpha_r, \mu_r, \sigma_r^2)$. Expressions (5) and (6) indicate how to sample from the posterior predictive distribution corresponding to the control and treatment groups after posterior simulation from $p(\boldsymbol{\eta}_r | \boldsymbol{Y}_r)$, $r = 1, 2$, is implemented. The Appendix provides details on how to obtain more general inference for functionals of the random mixtures $F(\cdot; G_r) = \int \text{Poisson}(\cdot|\theta)\, \mathrm{d}G_r(\theta)$, $r = 1, 2$ through sampling from the posterior of the mixing distributions G_r.

3.2 Stochastically ordered random-effects distributions

In certain applications, it might be useful to allow the control and treatment random effects distributions in (4) to be dependent. A special case of dependence for these distributions is induced by *stochastic ordering*, i.e., a prior assumption that either $G_1(\theta) \geq G_2(\theta)$, for all $\theta \in R$, or $G_1(\theta) \leq G_2(\theta)$, for all $\theta \in R$, denoted by $G_1 \leq_{st} G_2$ or $G_2 \leq_{st} G_1$, respectively. (See, e.g., Shaked and Shanthikumar 1994, for background on various types of probability orders.) Again, G_1 and G_2 are the random effects CDFs for the control and treatment groups, respectively. For instance, for the data example discussed in Section 1, the stochastic order restriction $G_2 \leq_{st} G_1$ provides a formal way to capture the assumption that the number of hospitalizations under the treatment cannot be larger than under the control. More typically, $G_1 \leq_{st} G_2$ would be the natural constraint, and, hence, our model development in this Section, as well as the examples of the simulation study in Section 4, are based on this assumption. A standard scenario where the $G_1 \leq_{st} G_2$ prior restriction might be useful involves medical applications with the Y_{ri} recording survival times (in discrete time) for a control group ($r = 1$) and a treatment group ($r = 2$), where it is known that treatment yields improvement (i.e., increase in mean survival time) and we are interested in assessing its extent.

From a practical perspective, for such applications the introduction of the stochastic order restriction in the prior will yield more accurate posterior predictive inference

(e.g., narrower posterior interval estimates). In our study, a prior over the space $\mathcal{P} = \{(G_1, G_2) \colon G_1 \leq_{st} G_2\}$ yields a second nonparametric model that can usefully be compared with (4).

A constructive approach to build such a prior is by considering the subspace $\mathcal{P}'$ of $\mathcal{P}$ defined by $\mathcal{P}' = \{(G_1, G_2) : G_1 = H_1, G_2 = H_1 H_2\}$, where H_1 and H_2 are CDFs on R, and placing independent DP priors on H_1 and H_2. Such a specification induces a nonparametric prior over $\mathcal{P}'$, and hence over (a subset of) $\mathcal{P}$, and allows for posterior inference based on extensions of MCMC methods for DP mixture models. This approach was developed in Gelfand and Kottas (2001) and Kottas and Gelfand (2001) (and also used in Kottas, Branco and Gelfand 2002). This earlier work was based on DP mixtures of normal distributions; here, we develop a different version for count data working with Poisson DP mixtures.

In particular, we assume $Y_{11}, \ldots, Y_{1n_1}$, given H_1, IID from the mixture $F(\cdot; H_1) = \int \text{Poisson}(\cdot; \theta_1)\, \mathrm{d}H_1(\theta_1)$, and $Y_{21}, \ldots, Y_{2n_2}$, given H_1 and H_2, IID from $F(\cdot; H_1, H_2) = \iint \text{Poisson}(\cdot; \max(\theta_1, \theta_2))\, \mathrm{d}H_1(\theta_1)\, \mathrm{d}H_2(\theta_2)$. Specifying independent DP priors for H_1 and H_2, we obtain the model

$$
\begin{array}{rcl}
Y_{1i}|\theta_{1i} & \overset{\text{indep}}{\sim} & \text{Poisson}(\theta_{1i}), \;\; i = 1, \ldots, n_1 \\
Y_{2k}|(\theta_{1,n_1+k}, \theta_{2k}) & \overset{\text{indep}}{\sim} & \text{Poisson}[\max(\theta_{1,n_1+k}, \theta_{2k})], \;\; k = 1, \ldots, n_2 \\
\theta_{1i}|H_1 & \overset{\text{IID}}{\sim} & H_1, \;\; i = 1, \ldots, n_1 + n_2 \\
\theta_{2k}|H_2 & \overset{\text{IID}}{\sim} & H_2, \;\; k = 1, \ldots, n_2 \\
H_r|(\alpha_r, \mu_r, \sigma_r^2) & \sim & \text{DP}(\alpha_r H_{r0}), \;\; r = 1, 2
\end{array}
\tag{7}
$$

where the DP base distributions $H_{r0} = \text{N}(\mu_r, \sigma_r^2)$, and the model is completed with hyperpriors for parameters α_r, μ_r, and σ_r^2, $r = 1, 2$. We refer to (7) as model $\mathcal{M}_3$.

The clustering of the latent variables $(\theta_{11}, \ldots, \theta_{1,n_1+n_2})$ is again represented by the number of clusters n_1^*, a vector of distinct cluster values $\boldsymbol{\theta}_1^* = (\theta_{1\ell}^*, \ell = 1, \ldots, n_1^*)$ and the indicator vector $\boldsymbol{s}_1 = (s_{11}, \ldots, s_{1,n_1+n_2})$, with $s_{1i} = \ell$ if and only if $\theta_{1i} = \theta_{1\ell}^*$ and $n_{1\ell} = |\{i : s_{1i} = \ell\}|$. The notation is analogous for the clustering structure of the $(\theta_{21}, \ldots, \theta_{2n_2})$. The predictive distribution for θ_1^{new} is

$$
p(\theta_1^{\text{new}}|\boldsymbol{\eta_1}) = \frac{\alpha_1}{\alpha_1 + n_1 + n_2} H_{10}(\theta_1^{\text{new}}|\mu_1, \sigma_1^2) + \frac{1}{\alpha_1 + n_1 + n_2} \sum_{\ell=1}^{n_1^*} n_{1\ell}\, \delta_{\theta_{1\ell}^*}(\theta_1^{\text{new}}),
$$

and similarly the predictive distribution for θ_2^{new} is

$$
p(\theta_2^{\text{new}}|\boldsymbol{\eta_2}) = \frac{\alpha_2}{\alpha_2 + n_2} H_{20}(\theta_2^{\text{new}}|\mu_2, \sigma_2^2) + \frac{1}{\alpha_2 + n_2} \sum_{\ell=1}^{n_2^*} n_{2\ell}\, \delta_{\theta_{2\ell}^*}(\theta_2^{\text{new}}),
$$

where $\boldsymbol{\eta_r} = (n_r^*, \boldsymbol{s}_r, \boldsymbol{\theta}_r^*, \alpha_r, \mu_r, \sigma_r^2)$, $r = 1, 2$. The posterior predictive distribution for a future Y_1^{new} is given by

$$p(Y_1^{\text{new}}|\boldsymbol{Y}_1, \boldsymbol{Y}_2) = \iint \text{Poisson}(Y_1^{\text{new}}|\theta_1^{\text{new}}) p(\theta_1^{\text{new}}|\boldsymbol{\eta}_1) p(\boldsymbol{\eta}_1|\boldsymbol{Y}_1, \boldsymbol{Y}_2),$$

and the posterior predictive distribution for Y_2^{new} is

$$\begin{aligned} p(Y_2^{\text{new}}|\boldsymbol{Y}_1, \boldsymbol{Y}_2) &= \int\int\int \text{Poisson}[Y_2^{\text{new}}|\max(\theta_1^{\text{new}}, \theta_2^{\text{new}})] \\ &\qquad p(\theta_1^{\text{new}}|\boldsymbol{\eta}_1) p(\theta_2^{\text{new}}|\boldsymbol{\eta}_2) p(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2|\boldsymbol{Y}_1, \boldsymbol{Y}_2). \end{aligned}$$

Details on more general types of inference, beyond the ones resulting from the posterior predictive distributions, are given in the Appendix.

4 Simulation study

We have conducted a simulation study fitting models $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ to four different data sets, each with control (D_{kC}) and treatment (D_{kT}) samples for $k = 1, \ldots, 4$. In each case we obtained posterior predictive distributions along with the posteriors of the random effects CDFs. The posterior sample size was equal to 1000 in all cases.

4.1 Data sets

All data sets were of size $n_1 = n_2 = 300$ (i.e., each of the control and treatment samples had 300 observations) and consisted of values drawn from Poisson[$\exp(\theta_{ri})$], $i = 1, \ldots, 300$, $r = 1, 2$, with the random-effects θ_{ri} generated as follows. For data set D_1 we used $\theta_{1i} \sim \text{N}(2.2, 0.65^2)$ and $\theta_{2i} \sim \text{N}(3.5, 0.5^2)$, i.e., with D_1 the assumptions of model $\mathcal{M}_1$ were satisfied. For data set D_2 the θ_{1i} were generated from a right skewed distribution (a four-component mixture of Normals) and the θ_{2i} were from a bimodal mixture of two Normals, $0.5\,\text{N}(3.3, 0.35^2) + 0.5\,\text{N}(5.8, 0.42^2)$. Data set D_3 was based on θ_{1i} values generated from $\text{N}(1.3, 0.5^2)$ and θ_{2i} values from $\text{N}(2.2, 0.5^2)$, i.e., the assumptions of $\mathcal{M}_1$ were again met but additionally the distribution of the θ_{1i} was stochastically smaller than the distribution of the θ_{2i}. Data set D_4 had θ_{1i} values drawn from $\text{N}(1.4, 0.4^2)$ and θ_{2i} values sampled from a bimodal mixture of two Normals, $0.5\,\text{N}(1.7, 0.37^2) + 0.5\,\text{N}(3.3, 0.52^2)$; again in this case the distribution of random effects for the control group was stochastically smaller than its treatment counterpart. Figure 1 summarizes all of the data sets and the corresponding random effects.

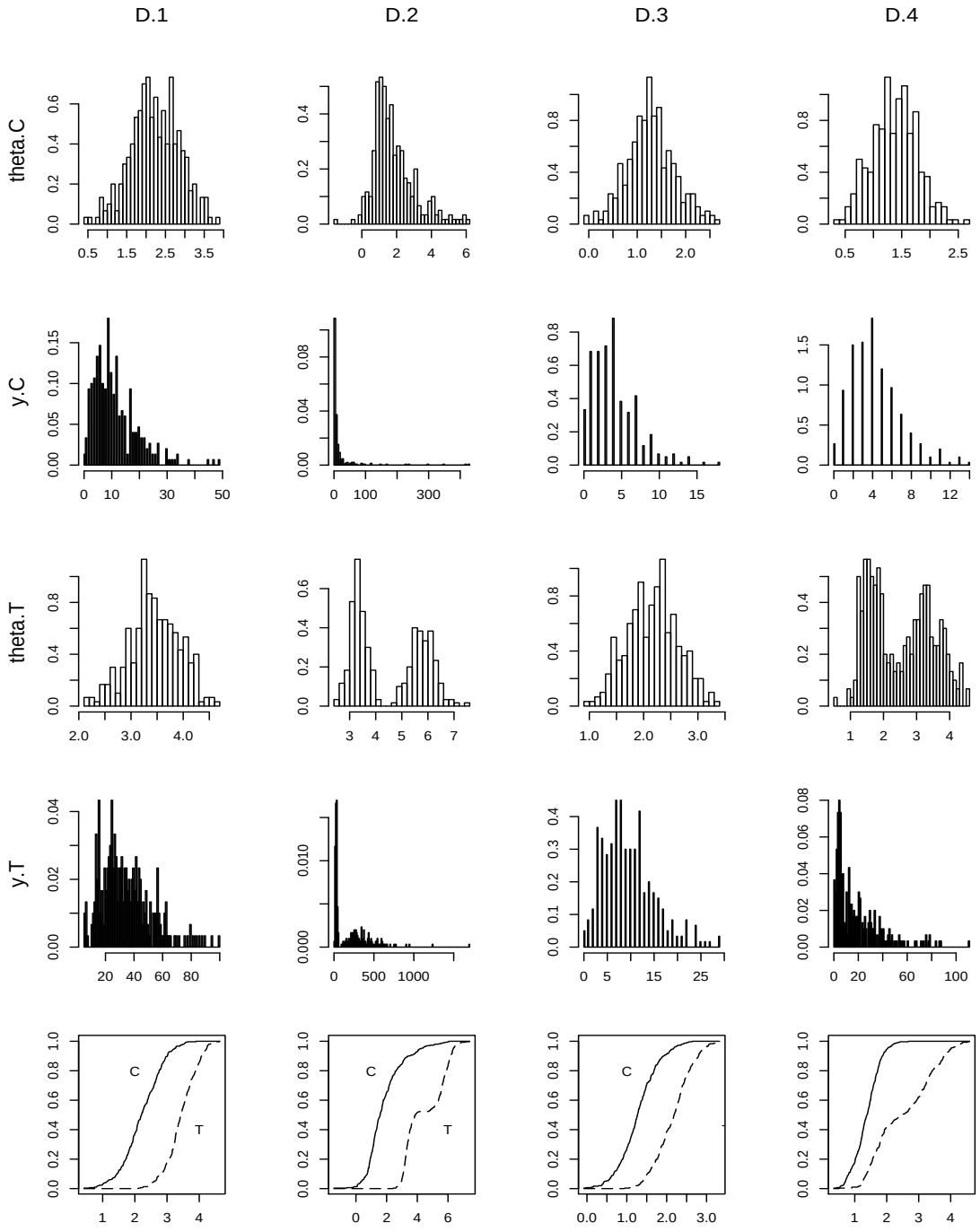

Figure 1: Data for the simulation study. Columns (from left to right) correspond to the four data sets (D_{kC}, D_{kT}), $k = 1, ..., 4$. Row 1 gives histograms of the latent variables θ_{1i} used to generate the count data Y_{1i} in row 2, with both of these rows applying to the simulated control groups; rows 3 and 4 have a similar structure for the treatment groups. The fifth row shows empirical CDFs of the latent variables (solid lines control, dashed lines treatment).

4.2 Prior specification

Model $\mathcal{M}_2$ reduces to the parametric random effects Poisson model $\mathcal{M}_1$ when $\alpha_r \to \infty$, which motivates specifying the same priors for the hyperparameters in these two models. We used normal priors for the μ_r with parameters determined based on (weak) information from the data (i.e., rough estimates for the center and range of the data). Inverse gamma priors with shape paramater equal to 2 (implying infinite variance) were used for the σ_r^2. Sensitivity analysis showed that values of the inverse gamma prior means around 1 resulted in stable inference and substantial learning about the σ_r^2. Regarding model $\mathcal{M}_3$, we used the same parametric prior forms for μ_r and σ_r^2, specifying their parameters through a straightforward extension of the approach described in Gelfand and Kottas (2001).

The priors for the α_r in models $\mathcal{M}_2$ and $\mathcal{M}_3$ were gamma distributions. As is well documented in the literature (e.g., Antoniak 1974; Escobar and West 1995; Liu 1996), in DP mixture models the DP precision parameter controls the clustering in the vector of mixing parameters. We conducted sensitivity analysis working with gamma priors that allowed for both small and large values for the α_r, and found that, with the sample sizes in this study, posterior predictive inference was very stable; moreover, there was visible posterior learning for the α_r.

4.3 Simulation results

We present inference results using posterior predictive distributions of future data values and posteriors for the random mixing distributions. Figure 2 gives an example of the posterior predictive distributions for the treatment part of the fourth data set under models $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$. It is clear that, while the prior predictive information is similarly diffuse for all three models, model $\mathcal{M}_1$ cannot adapt to the bimodality (without remodeling as, e.g., through a mixture of Gaussians on the latent scale), whereas the BNP models smoothly and automatically adapt to the data-generating mechanism.

It is also revealing to see how well the models recover the unknown mixing distributions. Figures 3 – 6 present posterior point and interval estimates for the CDFs of the mixing distributions G_r, $r = 1, 2$, for all data sets; in all of these figures, the long dotted lines track the true underlying CDF, the solid (blue) curves are the posterior mean estimates, and the short dotted lines give pointwise 90% posterior uncertainty bands. When the parametric model $\mathcal{M}_1$ is correct, as in Figure 3, it (naturally) yields narrower summaries of uncertainty than those generated by the BNP models, because it is a special case of BNP model $\mathcal{M}_2$ with stronger prior information ($\mathcal{M}_1$ assumes certainty about the form of the mixing distributions, whereas $\mathcal{M}_2$ embodies uncertainty about this structural

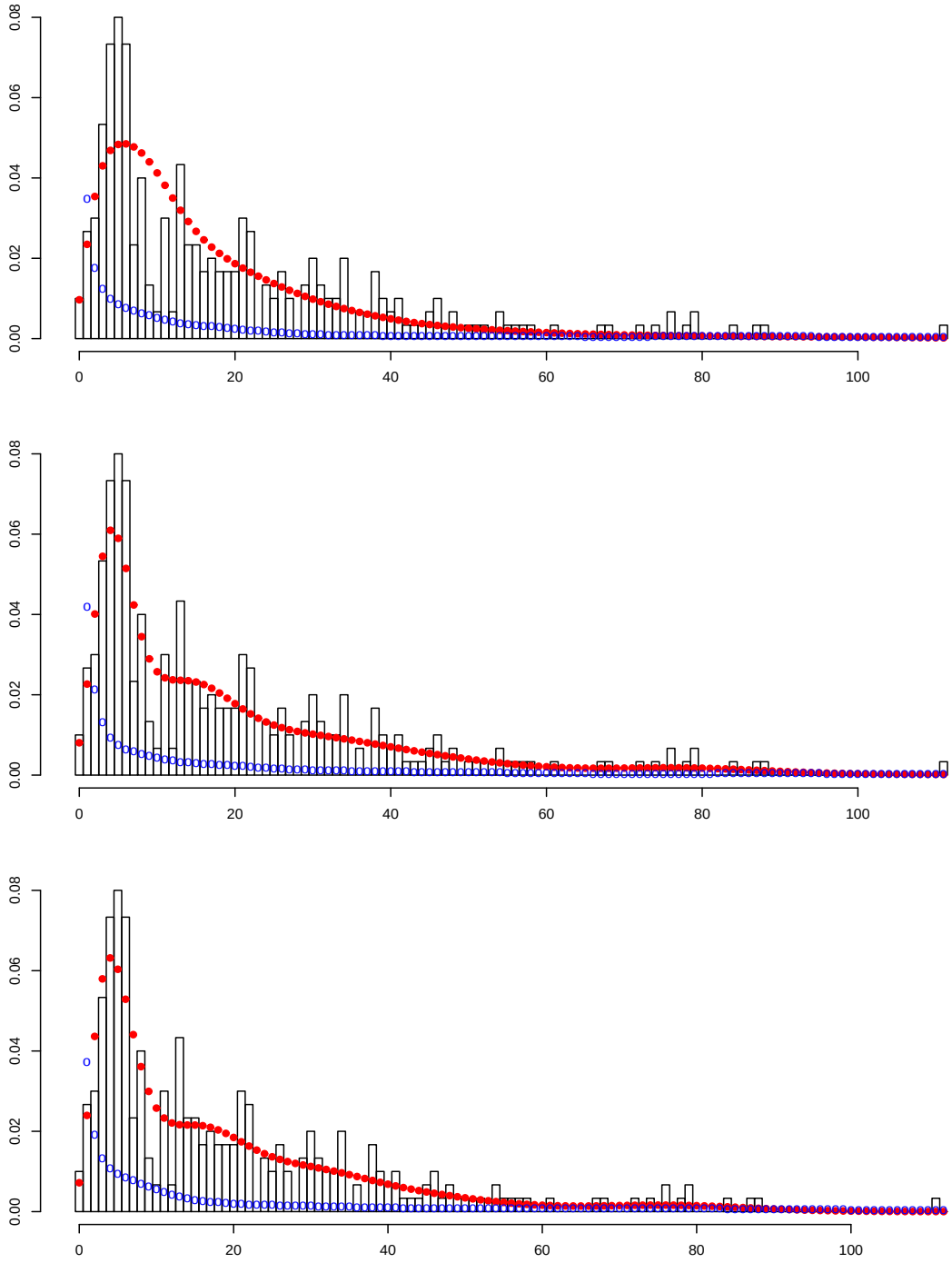

Figure 2: Simulation study. Prior (blue circles) and posterior (solid red circles) predictive distributions under models $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$ (top, middle and bottom panels, respectively) based on data set D_{4T}. In each panel, the histogram plots the simulated counts.

detail); however, when $\mathcal{M}_1$ is incorrect, as for example in Figures 4 and 6, it continues to yield narrower uncertainty bands that fail to include the truth, whereas BNP model $\mathcal{M}_2$ again adapts successfully to the skewed and bimodal data-generating mechanisms. Figures 5 and 6 demonstrate that the extra assumption of stochastic order, when true, yields narrower uncertainty bands (as it should). As a further illustration regarding the posterior $p(G_r|\boldsymbol{Y}_r)$, $r = 1, 2$, Figure 7 plots 30 realizations from these posteriors under model $\mathcal{M}_2$ and across all four data sets.

Another way to examine the performance of parametric and BNP models in our simulation study is to compare the predictive accuracy of models by contrasting the observed data values with the posterior predictive distributions under each model. The optimal way to carry out such a comparison (e.g., O'Hagan and Forster 2004; Draper and Krnjajić 2006) is with *log scoring* criteria, which reward a model based on the logarithm of the heights of the predictive distributions at the observed data points (so that higher log score values are preferable). With a data set $\boldsymbol{y} = (y_1, ..., y_n)$ consisting of n observations, and for model $\mathcal{M}$, two variants on this idea are worth examining: the cross-validation log score (e.g., Gelfand et al. 1992)

$$LS_{CV}(\mathcal{M}|\boldsymbol{y}) = \frac{1}{n}\sum_{j=1}^{n} \log[p(y_j|\boldsymbol{y}_{(-j)}, \mathcal{M})]$$

in which $\boldsymbol{y}_{(-j)}$ is the data vector $\boldsymbol{y}$ with observation y_j set aside, and the full-sample log score (e.g., Laud and Ibrahim 1995)

$$LS_{FS}(\mathcal{M}|\boldsymbol{y}) = \frac{1}{n}\sum_{j=1}^{n} \log[p(y_j|\boldsymbol{y}, \mathcal{M})]$$

which has both computational advantages and better small- and large-sample discriminating power (to identify good models) when compared with LS_{CV} (Draper and Krnjajić 2006). Table 2 summarizes log score results for parametric model $\mathcal{M}_1$ and BNP model $\mathcal{M}_2$ across the treatment and control samples in all four data sets illustrated in Figure 1. On the predictive data scale the two models fit about equally well when the parametric assumptions of $\mathcal{M}_1$ are met, but it is evident that $\mathcal{M}_2$ produces superior predictions when the latent variable distributions are skewed or bimodal.

5 A Bayesian nonparametric fixed-effects model for count data

Because the DP has (almost surely) discrete realizations, it is also possible to model the data in Table 1 by treating the underlying CDFs F_r of the control ($r = 1$) and the

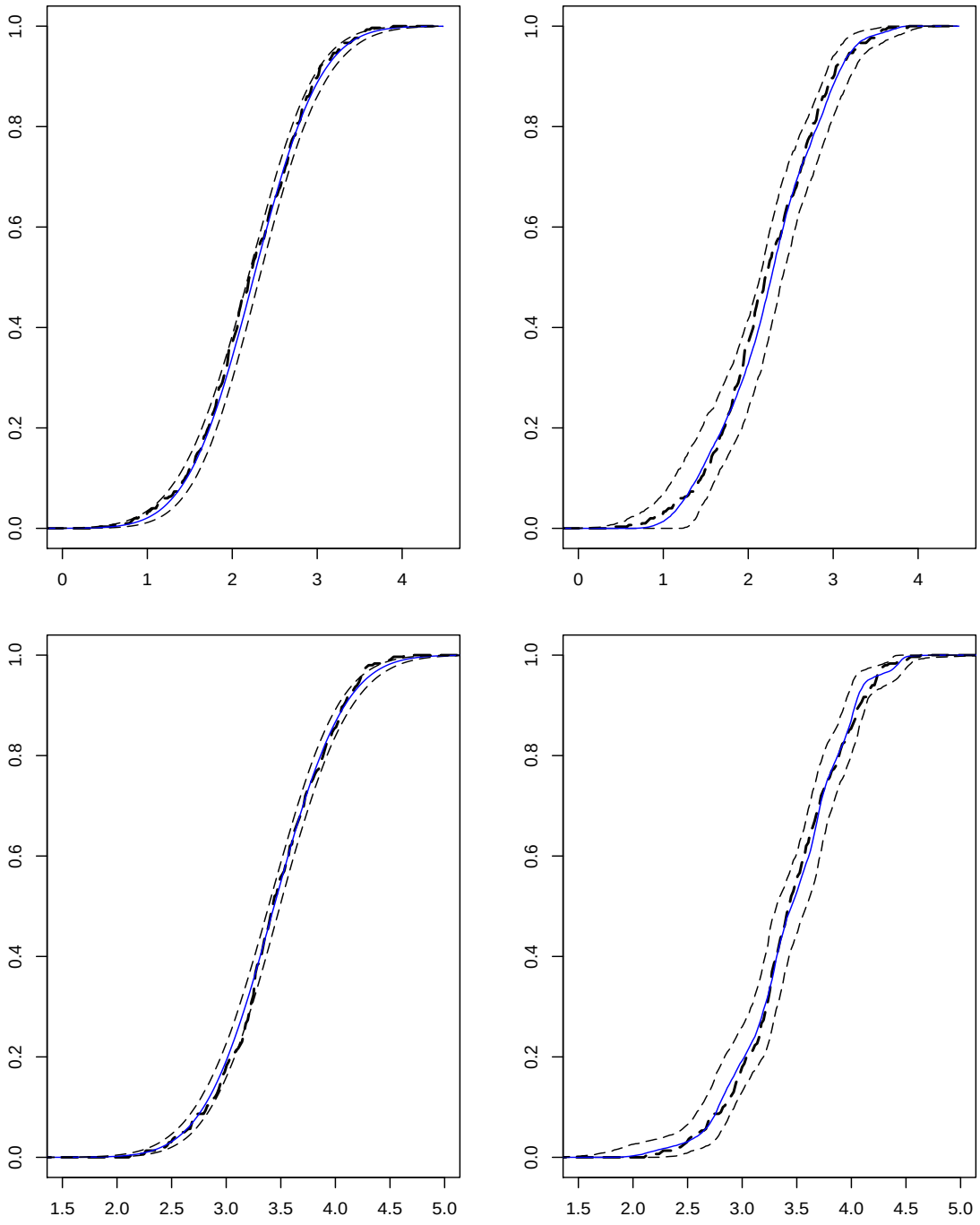

Figure 3: Simulation study. Posterior estimates of the mixing distributions under models $\mathcal{M}_1$ and $\mathcal{M}_2$ (left and right columns, respectively) and for data sets D_{1C} and D_{1T} (top and bottom rows, respectively).

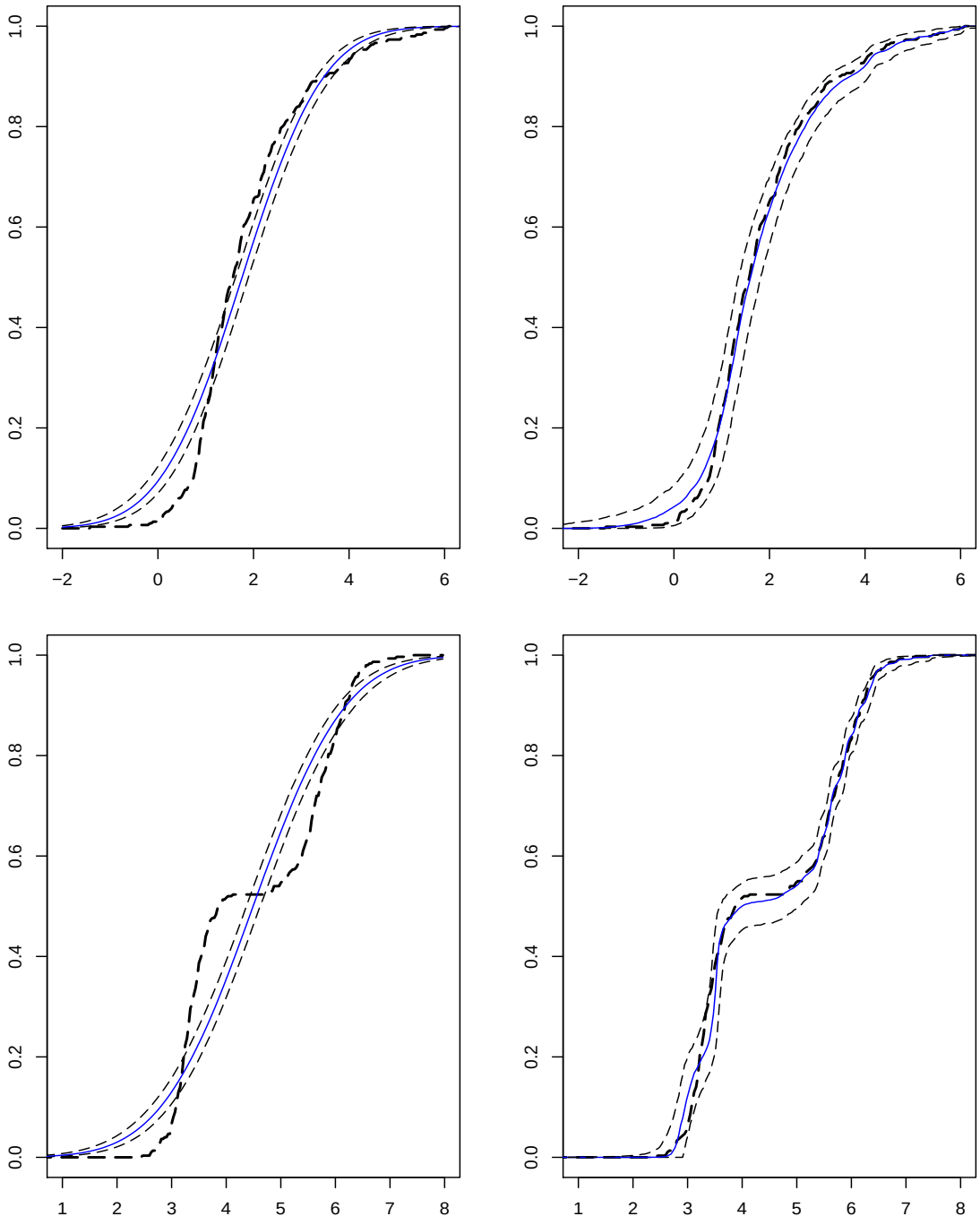

Figure 4: Simulation study. Posterior estimates of the mixing distributions under models $\mathcal{M}_1$ and $\mathcal{M}_2$ (left and right columns, respectively) and for data sets D_{2C} and D_{2T} (top and bottom rows, respectively).

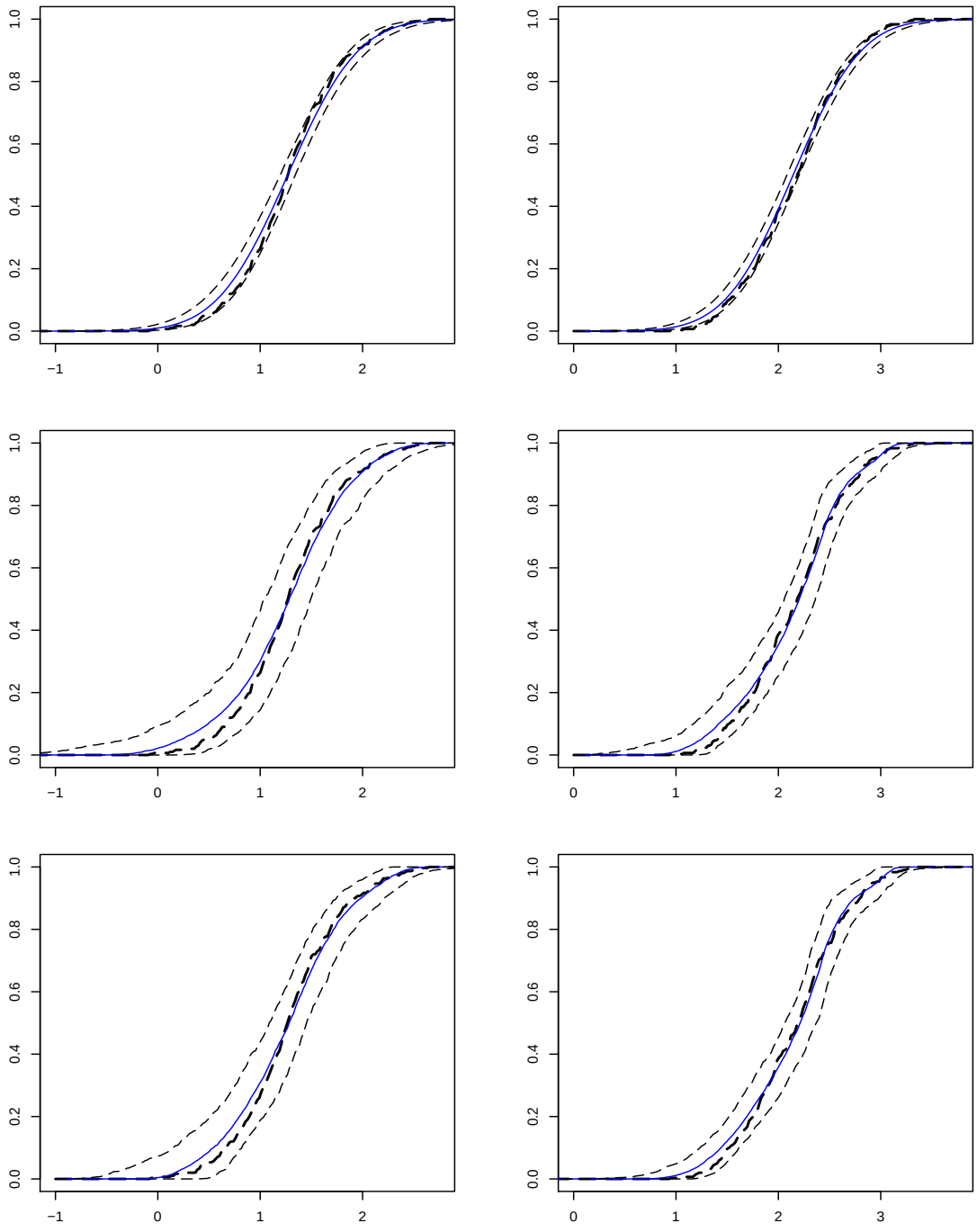

Figure 5: Simulation study. Posterior estimates of the mixing distributions for data sets D_{3C} and D_{3T} (left and right columns, respectively) under models $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$ (top, middle and bottom rows, respectively).

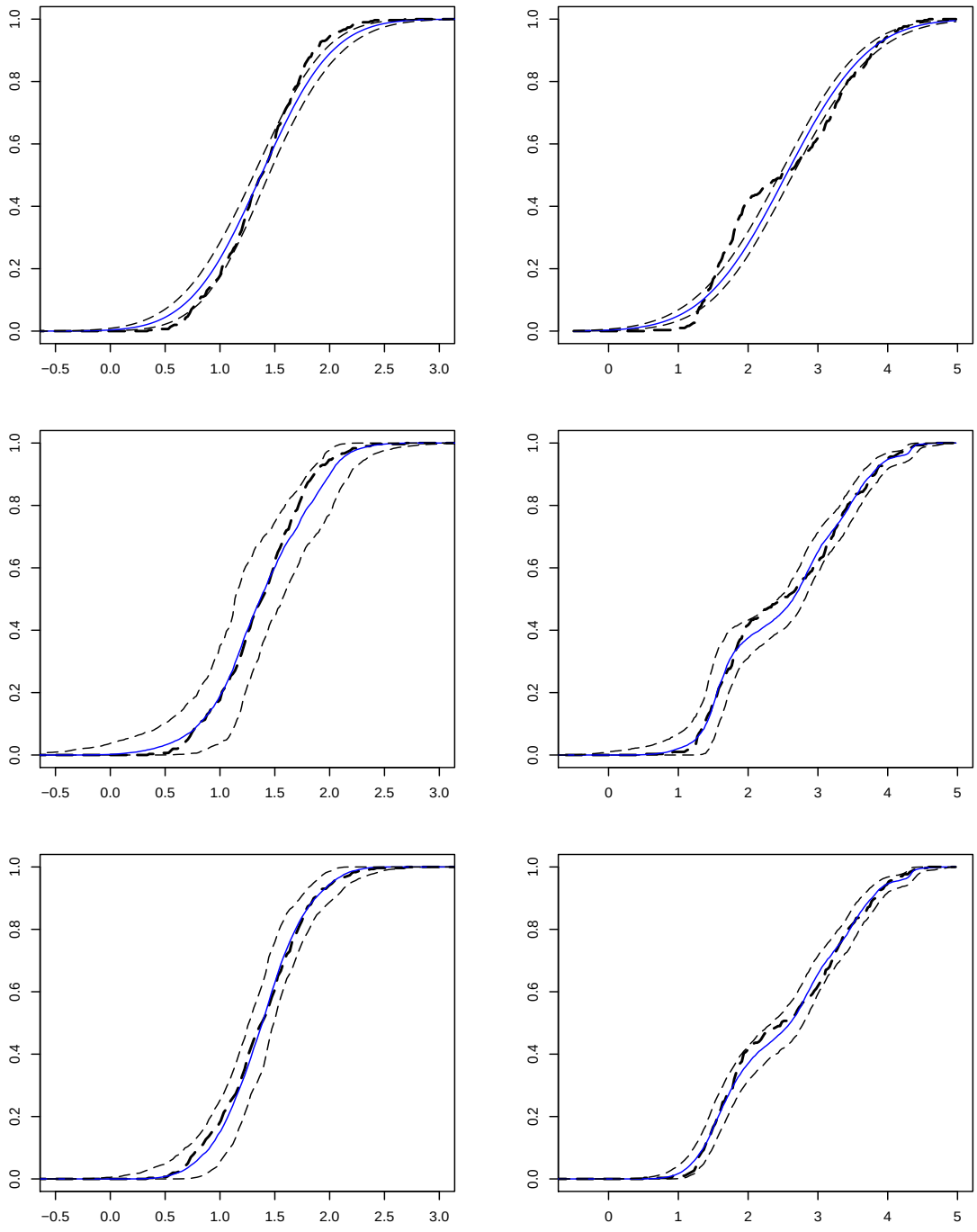

Figure 6: Simulation study. Posterior estimates of the mixing distributions for data sets D_{4C} and D_{4T} (left and right columns, respectively) under models $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$ (top, middle and bottom rows, respectively).

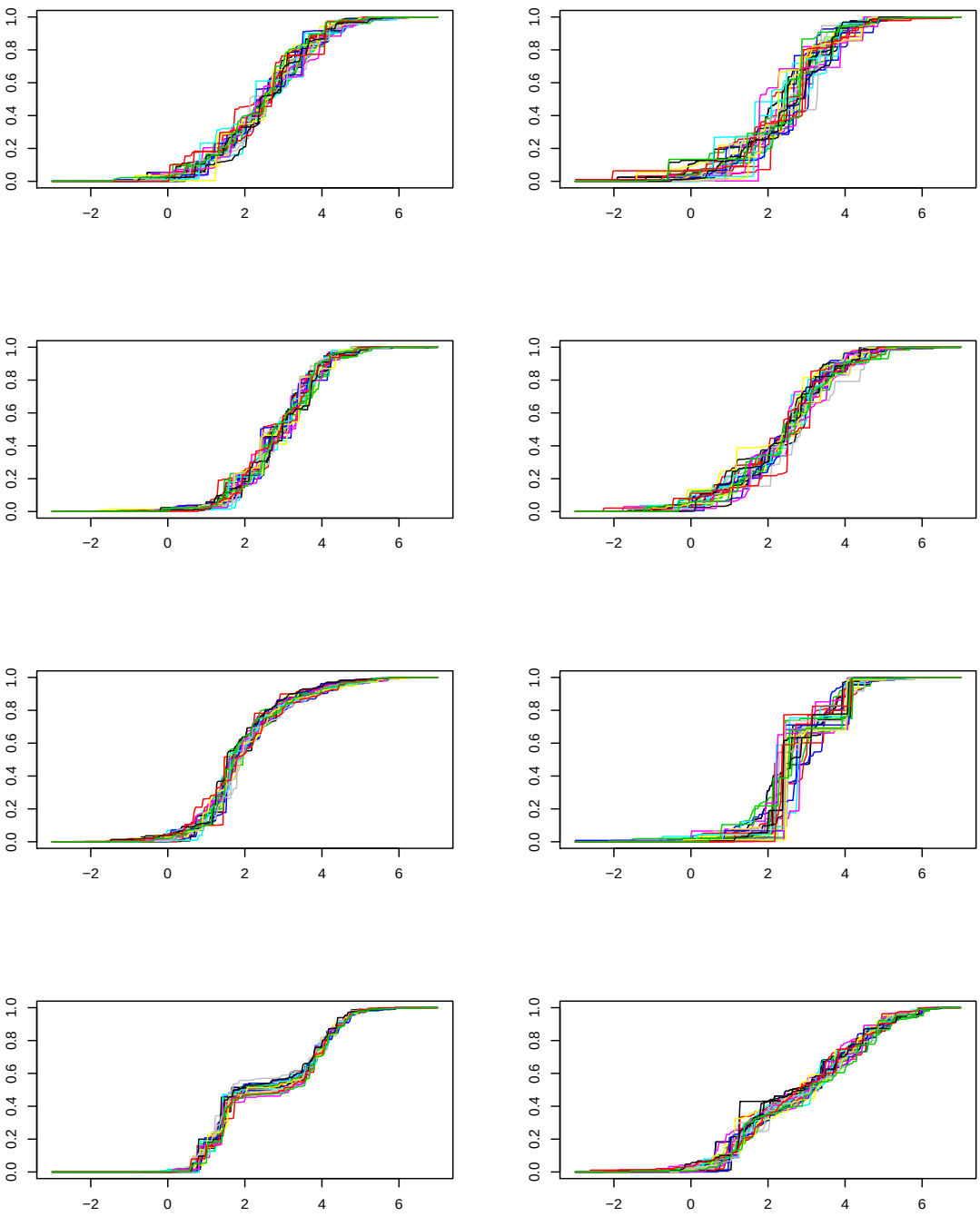

Figure 7: Simulation study. Posterior realizations for the random effects distributions G_1 and G_2 under model $\mathcal{M}_2$. Reading them from top to bottom, the left column panels correspond to data sets D_{1C}, D_{1T}, D_{2C}, D_{2T}; and the right column panels to data sets D_{3C}, D_{3T}, D_{4C}, D_{4T}.

Table 2: *Log score comparisons between parametric model $\mathcal{M}_1$ and nonparametric model $\mathcal{M}_2$ across the treatment and control samples in all four data sets (latent variable distributions: G = Gaussian, S = skewed, B = bimodal).*

	LS_{CV}							
	D_{1C} (G)	D_{1T} (G)	D_{2C} (S)	D_{2T} (B)	D_{3C} (G)	D_{3T} (G)	D_{4C} (G)	D_{4T} (B)
$\mathcal{M}_1$	–3.33	–4.23	–3.72	–6.22	–2.41	–3.07	–2.32	–3.99
$\mathcal{M}_2$	–3.35	–4.22	–3.52	–6.03	–2.42	–3.07	–2.31	–3.94
	LS_{FS}							
$\mathcal{M}_1$	–3.32	–4.22	–3.66	–6.19	–2.40	–3.07	–2.32	–3.97
$\mathcal{M}_2$	–3.32	–4.19	–3.46	–5.72	–2.41	–3.06	–2.31	–3.91

treatment ($r = 2$) groups as unknown and placing DP priors (centered, e.g., on the Poisson distribution) directly on the F_r. The result can be regarded as a BNP analogue of the parametric fixed-effects model $\mathcal{M}_0$ in (2). Specifically, for $r = 1, 2$, we consider the model (that will be referred to as model $\mathcal{M}_4$):

$$\begin{array}{rcl} Y_{ri}|F_r & \overset{\text{IID}}{\sim} & F_r, \;\; i = 1, ..., n_r \\ F_r|(\alpha_r, \theta_r) & \sim & \text{DP}[\alpha_r F_{r0}(\cdot|\theta_r)] \\ (\alpha_r, \theta_r) & \sim & p(\alpha_r)p(\theta_r), \end{array} \tag{8}$$

where the DP priors are independent, the base distribution F_{r0} is taken to be Poisson[$\exp(\theta_r)$], and $p(\alpha_r)$, $p(\theta_r)$ denote the hyperpriors. (We work with gamma priors for the α_r and with normal priors for the θ_r.) In the presence of covariate information, it is possible to extend model (8) to semiparametric settings (see, e.g., Carota and Parmigiani 2002).

To simplify notation, we drop the subscript r in the remainder of this Section. Because independent DP priors are used for F_r, $r = 1, 2$, the following results apply to both the treatment and control group data. Following Antoniak (1974, *Corollary* 3.2′), the joint posterior of F and (α, θ) can be expressed as

$$\begin{array}{rcl} p(F, \alpha, \theta|\boldsymbol{Y}) & = & p(F|\alpha, \theta, \boldsymbol{Y})p(\alpha, \theta|\boldsymbol{Y}) \\ & \propto & p(F|\alpha, \theta, \boldsymbol{Y})p(\alpha)p(\theta)L(\alpha, \theta; \boldsymbol{Y}). \end{array} \tag{9}$$

Here, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, and the distribution of $F|\alpha, \theta, \boldsymbol{Y}$ is a DP with precision parameter $\alpha + n$ and base CDF $F_0'(\cdot|\alpha, \theta, \boldsymbol{Y}) = \alpha(\alpha + n)^{-1}F_0(\cdot|\theta) + (\alpha + n)^{-1}\sum_{i=1}^{n} 1_{[Y_i,\infty)}(\cdot)$. Moreover, based on *Lemma 1* from Antoniak (1974), and given the discreteness of the

Table 3: *Log score comparisons between two parametric models ($\mathcal{M}_0$, fixed-effects Poisson; $\mathcal{M}_1$, random-effects Poisson) and two BNP models ($\mathcal{M}_2$, DP modeling on the random effects scale; $\mathcal{M}_4$, DP modeling on the data scale), applied to the IHGA data of Table 1.*

	LS_{CV}			
	$\mathcal{M}_0$	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_4$
Control	–1.404	–1.368	–1.347	–1.338
Treatment	–1.220	–1.211	–1.204	–1.206

	LS_{FS}			
Control	–1.398	–1.343	–1.342	–1.336
Treatment	–1.215	–1.199	–1.198	–1.205

base distribution F_0, the marginal likelihood for (α, θ) is given by

$$L(\alpha, \theta; \boldsymbol{Y}) = \frac{\alpha^{n^*}}{\alpha^{(n)}} \prod_{j=1}^{n^*} f_0(Y_j^*|\theta)\{\alpha f_0(Y_j^*|\theta) + 1\}^{n_j - 1},$$

where n^* is the number of distinct values Y_j^* in $\boldsymbol{Y}$, n_j is the size of data cluster j ($j = 1, ..., n^*$), $a^{(n)} = a(a+1)\cdots(a+n-1)$, and $f_0(x|\theta) = (x!)^{-1}\exp(\theta x - \exp(\theta))$.

To obtain samples from the marginal posterior $p(\alpha, \theta|\boldsymbol{Y}) \propto p(\alpha)p(\theta)L(\alpha, \theta; \boldsymbol{Y})$, we used a symmetric random-walk Metropolis algorithm on $(\log(\alpha), \theta)$. Next, based on (9), posterior realizations of F and any of its functionals (such as the mean functional) can be obtained by sampling from $p(F|\alpha, \theta, \boldsymbol{Y})$ for each posterior realization of (α, θ). This can be implemented using either the DP definition (Ferguson 1973) or the DP stick-breaking representation (Sethuraman 1994), both discussed in the Appendix. In particular, if we only seek the posterior of $F(y_0)$ for some specified non-negative integer y_0, expression (9) and the DP definition yield

$$p(F(y_0)|\boldsymbol{Y}) = \int\int p(F(y_0)|\alpha, \theta, \boldsymbol{Y}) p(\alpha, \theta|\boldsymbol{Y})\, \mathrm{d}\alpha\, \mathrm{d}\theta,$$

where $p(F(y_0)|\alpha, \theta, \boldsymbol{Y})$ is a Beta distribution with parameters $(\alpha + n)F_0'(y_0|\alpha, \theta, \boldsymbol{Y})$ and $(\alpha + n)(1 - F_0'(y_0|\alpha, \theta, \boldsymbol{Y}))$. In fact, this posterior suffices for the estimation of posterior predictive probabilities, $\Pr(Y = y_0|\boldsymbol{Y})$ (again, for any specified non-negative integer y_0), since we can write $\Pr(Y = y_0|\boldsymbol{Y}) = E\{\Pr(Y = y_0; F)|\boldsymbol{Y}\}$, and thus, $\Pr(Y = y_0|\boldsymbol{Y}) = E\{F(y_0) - F(y_0 - 1)|\boldsymbol{Y}\}$, for $y_0 = 1, 2, ...$, and $\Pr(Y = 0|\boldsymbol{Y}) = E\{F(1)|\boldsymbol{Y}\} - \Pr(Y = 1|\boldsymbol{Y})$.

6 Analysis of the IHGA data

We have fitted four models ($\mathcal{M}_0$, $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_4$) to the IHGA data set described in Section 1; Table 3 presents a log-score comparison of these models on the control and treatment samples. Using either LS_{CV} or LS_{FS}, essentially both of the BNP models are seen to be predictively superior to the parametric Poisson models. However, note that the differences of the BNP models from the random-effects parametric model are less pronounced than in their comparison with the fixed-effects parametric model (and, indeed, model $\mathcal{M}_1$ yields a larger LS_{FS} value than model $\mathcal{M}_4$ for the treatment data).

As noted in Section 1, the summary of the effect of the IHGA intervention of greatest policy relevance was the percentage decline in mean rate of hospitalization. Figure 8 presents posterior distributions of the mean functionals for the control and treatment groups (denoted by λ_1 and λ_2, respectively) and of the ratio $\eta = \lambda_2/\lambda_1$, obtained under models $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_4$. Recall from Section 2 that, under model $\mathcal{M}_1$, the mean functionals have a simple parametric form, $\lambda_r = \exp(\mu_r + 0.5\sigma_r^2)$, $r = 1, 2$. Although no such closed-form expressions exist for the BNP models, the posteriors of λ_1 and λ_2 under models $\mathcal{M}_2$ and $\mathcal{M}_4$ can be sampled as discussed in the Appendix and Section 5, respectively.

All three models agree inferentially that the IHGA intervention has been helpful in reducing hospitalization: point estimates of η range from 0.81 to 0.83 across the three models (i.e., IHGA is associated with an approximate 18% reduction in mean hospitalizations per two years in the population of patients judged exchangeable with those in the geriatric study), with a posterior standard deviation of about 0.09, and a posterior probability that $\eta < 1$ with values between 0.95 and 0.96 under the three models.

7 Discussion

In this paper we have contrasted two approaches to Bayesian model specification with count data — parametric (fixed-effects and random-effects Poisson modeling) and nonparametric (based on DP modeling, where the DP is placed either on the random effects or directly on the scale of the observations) — and we have demonstrated that the most natural parametric models are special cases of the BNP modeling approach we examine: in effect parametric modeling is just BNP modeling with stronger prior information (namely, that parametric models incorporate certainty about the precise form of sampling and mixing distributions, whereas BNP models treat these structural assumptions as uncertain). Contemporary computing resources and MCMC methods for integral ap-

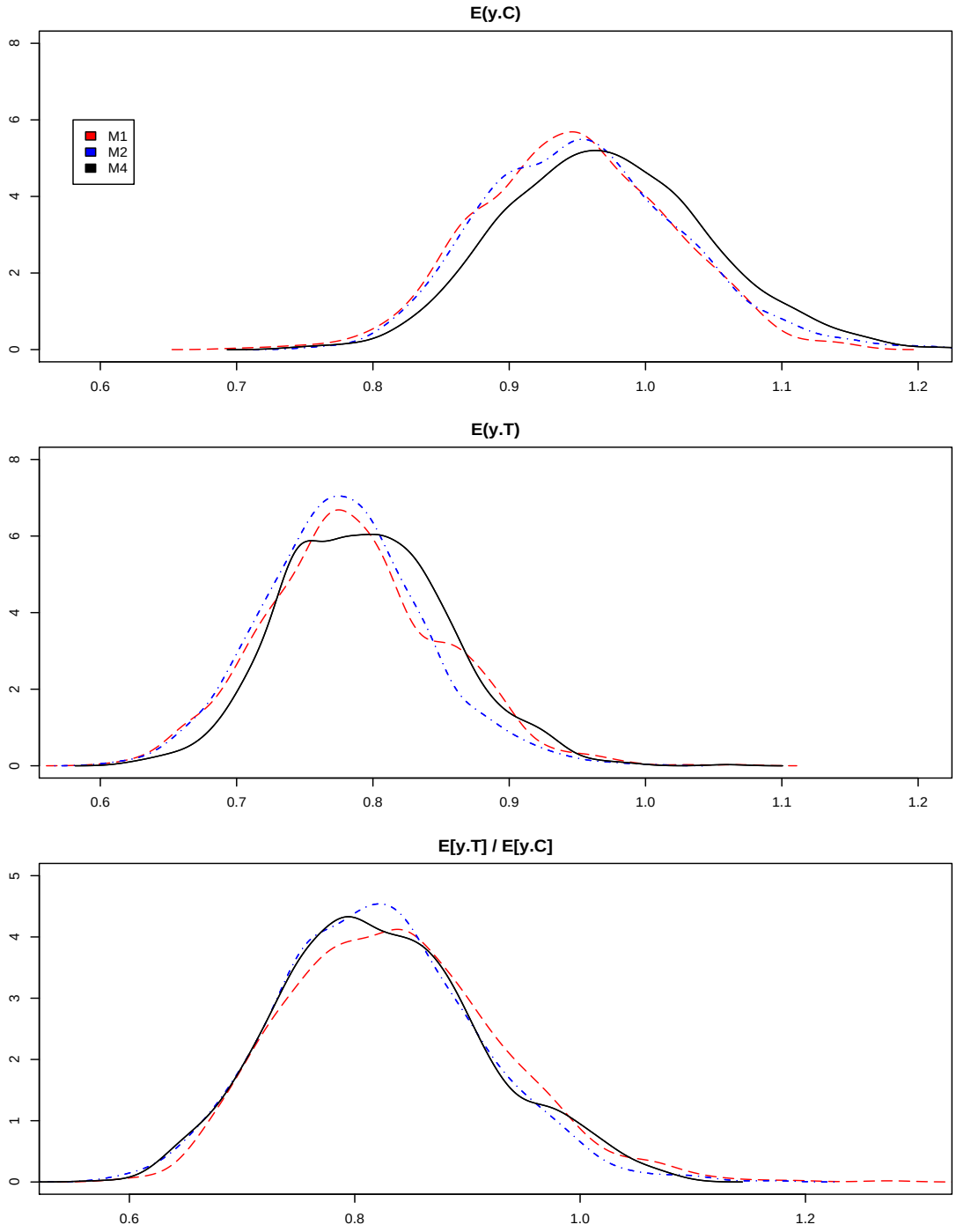

Figure 8: IHGA data. Posteriors of the mean functional λ_1 for the control group (top panel), the mean functional λ_2 for the treatment group (middle panel), and the policy-relevant ratio $\eta = \lambda_2/\lambda_1$ (bottom panel). Results are presented for models $\mathcal{M}_1$ (dashed lines), $\mathcal{M}_2$ (dot-dashed lines), and $\mathcal{M}_4$ (solid lines).

proximation combine to make both parametric and BNP approaches feasible and indeed increasingly straightforward.

In simulation studies like the one we have undertaken, when the stronger prior information inherent in parametric models is known to be true (by construction), the parametric models can yield narrower uncertainty bands than the BNP approach; this is simply an instance of the general phenomenon that stronger prior information may lead to less uncertainty. However, with a real data set obtained observationally rather than via simulation, the stronger prior information inherent in parametric models is no longer known to be true, and indeed when false parametric assumptions are made the resulting inferences may no longer be valid (Figures 4 and 6 in Section 4 exemplify this). BNP modeling is more validity-robust: in our simulation study, for example, DP mixture modeling was able to successfully recover a wide variety of underlying behavior (including skewness and bimodality in the random-effects distribution) without making strong assumptions about that behavior. This is an instance of the general adaptive character of BNP, about which useful results have been obtained; to informally restate a theorem of Walker, Damien and Lenk (2004) as an example, if $Y_i|F \overset{\text{IID}}{\sim} F$, for $i = 1, \ldots, n$, and a prior on F is used that places non-zero probability on all Kullback-Leibler neighborhoods of all densities (e.g., DP mixture models, typically, achieve this goal), then as $n \to \infty$, the posterior distribution $p(F|Y_1, \ldots, Y_n)$ will shrug off any incorrect details of prior specification (such as choosing a prior centering distribution F_0 that is far from F) and will fully adapt to the actual data-generating F. (Of course, if the conditional exchangeability judgments built into BNP models are false, even the BNP approach may not be validity-robust.)

As noted in Section 1, one may attempt within the realm of parametric modeling to recover inferential validity by a data-analytic process of

(1) parametric modeling,

(2) identifying defects in the model proposed in (1),

(3) parametric re-modeling,

(4) identifying defects in the model proposed in (3),

and so on, and indeed this approach is common in contemporary statistical work. But

(a) a price must be paid for identifying plausible structural assumptions in a data-driven manner, and

(b) it is substantially less common for investigators to demonstrate that they have paid the appropriate price for their data-driven searches through model space.

BNP modeling, by incorporating a more realistic initial uncertainty assessment about what is really known and unknown about the underlying structure of the "true" data-generating mechanism, provides an attractive alternative to parametric modeling, particularly in machine-learning settings in which the human intervention inherent in the data-analytic parametric approach is cumbersome or infeasible.

Appendix

To facilitate the presentation of our Bayesian nonparametric approaches to modeling count data, here we review basic definitions and results on Dirichlet processes (DPs) and DP mixtures. We also provide some details on the computational techniques for inference under the DP mixture models discussed in Section 3. The theory of DPs was established by Ferguson (1973, 1974), Blackwell (1973), Blackwell and MacQueen (1973) and Antoniak (1974) (building on work of Freedman 1963 and Fabius 1964). See Ghosh and Ramamoorthi (2003) for additional references on theoretical aspects of DP priors.

The Dirichlet process. The DP is a stochastic process with sample paths that can be interpreted as CDFs. Let Ω be a sample space and $\mathcal{F}$ a σ-field of subsets of Ω. According to the definition in Ferguson (1973), the DP is a stochastic process $Q = \{Q(\omega, A) \colon \omega \in \Omega, A \in \mathcal{F}\}$ with sample paths $\{Q_\omega(A) \equiv Q(\omega, A), \forall A \in \mathcal{F}\}$ that are probability measures on $(\Omega, \mathcal{F})$, such that, for any finite measurable partition $(A_1, \ldots, A_n)$ of Ω, the random vector $[Q(A_1), \ldots, Q(A_n)]$ has a Dirichlet distribution with parameters $[\alpha Q_0(A_1), \ldots, \alpha Q_0(A_n)]$. Here, α is a positive scalar parameter and Q_0 a specified probability measure on $(\Omega, \mathcal{F})$. Hence, $Q_0(A)$ (a constant) and $Q(A)$ (a random variable) denote the probability of event A under Q_0 and Q, respectively. Therefore, Q can be viewed as a random probability measure on $(\Omega, \mathcal{F})$. It is easy to show from basic properties of the Dirichlet distribution that for any $A \in \Omega$, $E[Q(A)] = Q_0(A)$ and $V[Q(A)] = Q_0(A)[1 - Q_0(A)]/(\alpha + 1)$. Hence, Q_0 is the center of the DP and α can be interpreted as a precision parameter.

A DP is, formally, defined on the space of probability measures but we often use the term distribution or CDF instead. For example, when $\Omega = R$ and $A = (-\infty, x), x \in R$, then $Q(A) = G(x)$ has a Beta distribution with parameters $\alpha\, G_0(x)$ and $\alpha[1 - G_0(x)]$ and, thus, $E[G(x)] = G_0(x)$ and $V[G(x)] = G_0(x)[1 - G_0(x)]/(\alpha + 1)$, where G is a random CDF and G_0 is a specified CDF on R. For larger values of α, a realization G from the DP is expected to be closer to the centering (or base) distribution G_0. We will 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 $G_0(\cdot) \equiv G_0(\cdot|\boldsymbol{\psi})$.

An important alternative definition of the DP was given by Sethuraman (1994) (see also Sethuraman and Tiwari 1982). This is a constructive definition that represents DP realizations as countable mixtures of point masses. Specifically, let $\{z_k, k = 1, 2, \dots\}$ and $\{\vartheta_j, j = 1, 2, \dots\}$ be independent sequences of IID random variables with $z_k \sim \text{Beta}(1, \alpha)$ and $\vartheta_j \sim \text{G}_0$, and define the weights through the following *stick-breaking* procedure: $w_1 = z_1$, $w_i = z_i \prod_{k=1}^{i-1}(1 - z_k), i = 2, 3, \dots$ Then a realization G from $\text{DP}(\alpha G_0)$ is (almost surely) of the form

$$G(\cdot) = \sum_{i=1}^{\infty} w_i \, \delta_{\vartheta_i}(\cdot), \tag{10}$$

where $\delta_x(\cdot)$ denotes a point mass at x. 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 IID 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., Hjort 2000; Ishwaran and Zarepour 2000; Ongaro and Cattaneo 2004), versions of the DP that enable full posterior inference (Muliere and Tardella 1998; Gelfand and Kottas 2002), and prior models for dependent distributions (e.g., MacEachern 2000; De Iorio et al. 2004; Gelfand, Kottas and MacEachern 2005; Griffin and Steel 2006; Teh et al. 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

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

with the analogous notation for the random mixture density $f(\cdot; G)$. The kernel can be chosen to be a continuous distribution (thus overcoming the almost sure discreteness of the DP) or a discrete distribution as in the DP mixtures developed in Section 3, where $K(\cdot|\boldsymbol{\theta})$ is Poisson with $\boldsymbol{\theta} \equiv \theta \in R$.

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, IID 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:

$$\begin{array}{rcl} Y_i|\boldsymbol{\theta}_i & \overset{\text{indep}}{\sim} & K(\cdot|\boldsymbol{\theta}_i), \;\; i = 1, ..., n \\ \boldsymbol{\theta}_i|G & \overset{\text{IID}}{\sim} & G, \;\; i = 1, ..., n \\ G \mid \alpha, \boldsymbol{\psi} & \sim & \text{DP}(\alpha G_0(\cdot|\boldsymbol{\psi})). \end{array} \tag{11}$$

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 (11) where G is marginalized over its DP prior,

$$\begin{array}{rcl} Y_i|\boldsymbol{\theta}_i & \overset{\text{indep}}{\sim} & K(\cdot|\boldsymbol{\theta}_i), \;\; i = 1, ..., n \\ (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)|(\alpha, \boldsymbol{\psi}) & \sim & p(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n|\alpha, \boldsymbol{\psi}) \end{array} \tag{12}$$

where

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

The structure of this version of the model is central to the development of most of the simulation-based model fitting methods for DP mixtures.

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. This class of BNP 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; Jain and Neal 2004); see Liu (1996), MacEachern, Clyde and Liu (1999), and Blei and Jordan (2006) for alternative approaches.

Working with model (12), any of these posterior simulation methods can be utilized to obtain samples from the marginal posterior $p(\boldsymbol{\theta}, \alpha, \boldsymbol{\psi}|\boldsymbol{Y})$, and these samples can be used to estimate posterior predictive densities (as shown in, e.g., Escobar and West 1995). Posterior predictive inference is, typically, sufficient in, say, density estimation applications or semiparametric settings where inference for G is of less importance. However, in

fully nonparametric settings (such as the one in this paper), full posterior inference for the mixing distribution G, and for the mixture $F(\cdot; G)$ and any of its functionals, is essential. Note that, if Y^{new} denotes a future observable (under model (11)), its posterior predictive distribution (density) evaluated, say, at point z, is simply the posterior expectation of the CDF (PDF) functional (at point z) of $F(\cdot; G)$, e.g., $p(z|\boldsymbol{Y}) = E\{f(z; G)|\boldsymbol{Y}\}$ for the posterior predictive density. We describe next how to obtain general inferences for the two DP mixture models discussed in Section 3, using the approach in Gelfand and Kottas (2002) and Kottas (2006).

General inference for the Poisson DP mixtures. Consider first the DP mixture model presented in Section 3.1. Because the DP priors for G_1 and G_2 are independent, the approach is the same for the control group ($r = 1$) and treatment group ($r = 2$) mixture models, $F(\cdot; G_r) = \int \text{Poisson}(\cdot|\theta)\, \text{d}G_r(\theta)$, and we thus drop the subscript r in the following description. Based, again, on Antoniak (1974), the joint posterior for the random effects distribution G, the vector of random effects $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$, and the DP hyperparameters (α, μ, σ^2) is given by

$$p(G, \boldsymbol{\theta}, \alpha, \mu, \sigma^2|\boldsymbol{Y}) = p(G|\boldsymbol{\theta}, \alpha, \mu, \sigma^2) p(\boldsymbol{\theta}, \alpha, \mu, \sigma^2|\boldsymbol{Y}) \tag{13}$$

where $\boldsymbol{Y} = (Y_1, ..., Y_n)$, and $p(G|\boldsymbol{\theta}, \alpha, \mu, \sigma^2)$ indicates a DP with precision parameter $\tilde{\alpha} = \alpha + n$ and base CDF

$$\tilde{G}_0(t) \equiv \tilde{G}_0(t|\boldsymbol{\theta}, \alpha, \mu, \sigma^2) = \frac{\alpha}{\alpha + n} G_0(t|\mu, \sigma^2) + \frac{1}{\alpha + n} \sum_{i=1}^{n} 1_{[\theta_i, \infty)}(t).$$

As discussed in Section 3.1, a posterior simulation algorithm from Neal (2000) was used to obtain samples $(\boldsymbol{\theta}_b, \alpha_b, \mu_b, \sigma_b^2 : b = 1, ..., B)$ from the marginal posterior $p(\boldsymbol{\theta}, \alpha, \mu, \sigma^2|\boldsymbol{Y})$. Therefore, based on (13), a sample from the entire posterior can be obtained by drawing G_b from $p(G|\boldsymbol{\theta}_b, \alpha_b, \mu_b, \sigma_b^2)$ for $b = 1, ..., B$. To this end, the standard approach (discussed in Gelfand and Kottas 2002), involves a truncation approximation to the DP stick-breaking representation in (10), and indeed this is the most general approach for multivariate mixing distributions G. However, for our model, G is a CDF on R and thus we can use a computationally simpler technique. Specifically, consider a grid of points $t_1 < t_2 < \cdots < t_L$ on the real line. Then, we can draw from the posterior distribution of $\{G(t_1), \ldots, G(t_L)\}$ noting that, based on the original DP definition, $\{G(t_1), G(t_2) - G(t_1), ..., G(t_L) - G(t_{L-1}), 1 - G(t_L)\}$ has a Dirichlet distribution with parameters

$$\left\{\tilde{\alpha}\tilde{G}_0(t_1), \tilde{\alpha}[\tilde{G}_0(t_2) - \tilde{G}_0(t_1)], \ldots, \tilde{\alpha}[\tilde{G}_0(t_L) - \tilde{G}_0(t_{L-1})], \tilde{\alpha}[1 - \tilde{G}_0(t_L)]\right\}.$$

Hence, by sampling from the ordered Dirichlet distribution above for each posterior sample $(\boldsymbol{\theta}_b, \alpha_b, \mu_b, \sigma_b^2)$, we obtain, up to the finite grid approximation, B posterior realizations $\{G_b(t_1), \ldots, G_b(t_L)\}$ for the CDF $\{G(t) : t \in R\}$. These posterior samples were used to provide, e.g., the estimates in Figures 3 – 6. Moreover, they enable general inference for the mixture CDF $F(y; G)$ and for any functional that emerges from the mixture distribution $F(\cdot; G)$, for instance, the mean functional, $\lambda = \int y \, \mathrm{d}F(y; G) = \sum_{y=0}^{\infty} y\{\int \text{Poisson}(y|\theta) \, \mathrm{d}G(\theta)\} = \int \exp(\theta) \, \mathrm{d}G(\theta)$, which was used in the illustrative analysis of the IHGA data in Section 6.

Turning to the stochastically ordered Poisson DP mixture model of Section 3.2, the full posterior corresponding to (7) is given by

$$p(H_1, H_2, \boldsymbol{\eta}_1, \boldsymbol{\eta}_2 | \boldsymbol{Y}_1, \boldsymbol{Y}_2) = p(H_1|\boldsymbol{\eta}_1) p(H_2|\boldsymbol{\eta}_2) p(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 | \boldsymbol{Y}_1, \boldsymbol{Y}_2)$$

where $\boldsymbol{\eta}_r = (\boldsymbol{\theta}_r, \alpha_r, \mu_r, \sigma_r^2)$, $r = 1, 2$, with $\boldsymbol{\theta}_1 = (\theta_{11}, ..., \theta_{1,n_1+n_2})$ and $\boldsymbol{\theta}_2 = (\theta_{21}, ..., \theta_{2n_2})$, and $p(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 | \boldsymbol{Y}_1, \boldsymbol{Y}_2)$ is the marginal posterior resulting from model (7) after marginalizing H_1 and H_2 over their DP priors. We used an extension of the MCMC algorithm in Gelfand and Kottas (2001) to sample from $p(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 | \boldsymbol{Y}_1, \boldsymbol{Y}_2)$. Again, the samples from this posterior, combined with sampling from two DPs, yield the full posterior of the model, since the conditional posterior distributions $p(H_r|\boldsymbol{\eta}_r)$, $r = 1, 2$, are DPs. Here, $p(H_1|\boldsymbol{\eta}_1)$ is a DP with precision parameter $\alpha_1 + n_1 + n_2$ and base distribution with point masses $(\alpha_1 + n_1 + n_2)^{-1}$ at θ_{1i}, $i = 1, ..., n_1 + n_2$, and continuous mass $\alpha_1(\alpha_1 + n_1 + n_2)^{-1}$ on H_{10}; the precision parameter of $p(H_2|\boldsymbol{\eta}_2)$ is given by $\alpha_2 + n_2$ and its base distribution has point masses $(\alpha_2 + n_2)^{-1}$ at θ_{2k}, $k = 1, ..., n_2$, and continuous mass $\alpha_2(\alpha_2 + n_2)^{-1}$ on H_{20}. Posterior realizations of H_1 and H_2 yield directly the posteriors of the mixing distributions $G_1 = H_1$ and $G_2 = H_1 H_2$, including the estimates of the CDFs $\{G_1(t) : t \in R\}$ and $\{G_2(t) : t \in R\}$ (again, up to a finite grid approximation) plotted in Figures 5 and 6.

References

Antoniak CE (1974). Mixtures of Dirichlet processes with applications to nonparametric problems. *Annals of Statistics*, **2**, 1152–1174.

Blackwell D (1973). Discreteness of Ferguson selections. *Annals of Statistics*, **1**, 356–358.

Blackwell D, MacQueen JB (1973). Ferguson distributions via Pólya urn schemes. *Annals of Statistics*, **1**, 353–355.

Blei DM, Jordan MI (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, **1**, 121–144.

Brunner LJ, Lo AY (1989). Bayes methods for a symmetric unimodal density and its mode. *The Annals of Statistics*, **17**, 1550–1566.

Bush CA, MacEachern SN (1996). A semiparametric Bayesian model for randomized block designs. *Biometrika*, **83**, 275–285.

Carota C, Parmigiani G (2002). Semiparametric regression for count data. *Biometrika*, **89**, 265–281.

De Iorio M, Müller P, Rosner GL, MacEachern SN (2004). An ANOVA model for dependent random measures. *Journal of the American Statistical Association*, **99**, 205–215.

Draper D (1995). Assessment and propagation of model uncertainty (with discussion and rejoinder). *Journal of the Royal Statistical Society, Series B*, **57**, 45–97.

Draper D, Krnjajić M (2006). Bayesian model specification. Submitted.

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

Fabius J (1964). Asymptotic behavior of Bayes estimates. *Annals of Mathematical Statistics*, **35**, 846–856.

Ferguson TS (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1**, 209–230.

Ferguson TS (1974). Prior distributions on spaces of probability measures. *Annals of Statistics*, **2**, 615-629.

Ferguson TS (1983). Bayesian density estimation by mixtures of normal distributions. In *Recent Advances in Statistics* (Rizvi MH, Rustagi JS, Siegmund D, editors), New York: Academic Press, pp. 287–302.

Freedman DA (1963). On the asymptotic behavior of Bayes estimates in the discrete case. *Annals of Mathematical Statistics*, **34**, 1386–1403.

Gelfand AE, Kottas A (2001). Nonparametric Bayesian modeling for stochastic order. *Annals of the Institute of Statistical Mathematics*, **53**, 865–876.

Gelfand AE, 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.

Gelfand AE, Dey DK, Chang H (1992). Model determination using predictive distributions, with implementation via sampling-based methods (with discussion). In *Bayesian Statistics 4* (Bernardo JM, Berger JO, Dawid AP, Smith AFM, editors), Oxford: Oxford University Press, pp. 147–167.

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

Ghosh JK, Ramamoorthi RV (2003). *Bayesian Nonparametrics*. New York: Springer.

Griffin JE, Steel MFJ (2006). Order-based dependent Dirichlet processes. *Journal of the American statistical Association*, **101**, 179–194.

Hanson T, Branscum A, Johnson W (2005). Bayesian nonparametric modeling and data analysis: An introduction. In *Handbook of Statistics, volume 25: Bayesian Thinking, Modeling and Computation* (Dey DK and Rao CR, editors), Amsterdam: Elsevier, pp. 245–278.

Hendriksen C, Lund E, Stromgard E (1984). Consequences of assessment and intervention among elderly people: a three year randomized controlled trial. *British Medical Journal*, **289**, 1522–1524.

Hjort NL (2000). Bayesian analysis for a generalised Dirichlet process prior. Statistical Research Report, Department of Mathematics, University of Oslo.

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

Ishwaran H, Zarepour M (2000). Markov chain Monte Carlo in approximate Dirichlet and Beta two-parameter process hierarchical models. *Biometrika*, **87**, 371–390.

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

Kottas A (2006). Nonparametric Bayesian survival analysis using mixtures of Weibull distributions. *Journal of Statistical Planning and Inference*, **136**, 578–596.

Kottas A, Gelfand AE (2001). Modeling variability order: a semiparametric Bayesian approach. *Methodology and Computing in Applied Probability*, **3**, 427–442.

Kottas A, Branco MD, Gelfand AE (2002). A nonparametric Bayesian modeling approach for cytogenetic dosimetry. *Biometrics*, **58**, 593–600.

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

Laud P, Ibrahim J (1995). Predictive model selection. *Journal of the Royal Statistical Society B*, **57**, 247–262.

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

Lo AY (1984). On a class of Bayesian nonparametric estimates: density estimates. *Annals of Statistics*, **12**, 351–357.

MacEachern SN (2000). Dependent Dirichlet processes. Technical Report, Department of Statistics, The Ohio State University.

MacEachern SN, Müller P (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, **7**, 223–238.

MacEachern SN, Clyde M, Liu JS (1999). Sequential importance sampling for nonparametric Bayes models: The next generation. *The Canadian Journal of Statistics*, **27**, 251–267.

Muliere P, Tardella L (1998). Approximating distributions of random functionals of Ferguson-Dirichlet priors. *The Canadian Journal of Statistics*, **26**, 283–297.

Müller P, Quintana FA (2004). Nonparametric Bayesian data analysis. *Statistical Science*, **19**, 95–110.

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

O'Hagan A, Forster J (2004). *Bayesian Inference*, second edition. London: Arnold.

Ongaro A, Cattaneo C (2004). Discrete random probability measures: a general framework for nonparametric Bayesian inference. *Statistics & Probability Letters*, **67**, 33–45.

Paddock SM, Ridgeway G, Lin R, Louis TA (2006). Flexible distributions for triple-goal estimates in two-stage hierarchical models. *Computational Statistics & Data Analysis*, **50**, 3243–3262.

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

Sethuraman J, Tiwari R (1982). Convergence of Dirichlet measures and the interpretation of their parameter. In *Proceedings of the Third Purdue Symposium on Statistical Decision Theory and Related Topics*, Gupta SS, Berger J (editors), New York: Academic Press, 305–315.

Shaked M, Shanthikumar JG (1994). *Stochastic Orders and Their Applications*. Boston: Academic Press.

Teh YW, Jordan MI, Beal MJ, Blei DM (2006). Hierarchical Dirichlet processes. To appear in the *Journal of the American Statistical Association*.

Walker S, Damien P, Lenk P (2004). On priors with a Kullback-Leibler property. *Journal of the American Statistical Association*, **99**, 404–408.

Walker S, Damien P, Laud P, Smith AFM (1999). Bayesian nonparametric inference for random distributions and related functions (with discussion). *Journal of the Royal Statistical Society Series B*, **61**, 485–527.