Parametric and nonparametric Bayesian model specification: a case study

Milovan Krnjajić, David Draper†, and Athanasios Kottas

University of California, Santa Cruz, 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 (DP) mixture and mixtures of DP 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, fixed-effects Poisson model, MCMC, 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 (C) group, who received standard health care, and 285 to a treatment (T) 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) ≈ 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 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; the first choice might well be a fixed-effects model (in the absence of strong prior information about the underlying hospitalization rates in the C and T groups) of the form

\[ (\lambda^C, \lambda^T) \sim \text{diffuse}, \]
\[ (C_i|\lambda^C) \overset{\text{iid}}{\sim} \text{Poisson}(\lambda^C), \quad (T_j|\lambda^T) \overset{\text{iid}}{\sim} \text{Poisson}(\lambda^T), \]

\[ (1) \]

†Address for correspondence: Department of Applied Mathematics and Statistics, Baskin School of Engineering, University of California, 1156 High Street, Santa Cruz CA 95064 USA (email draper@ams.ucsc.edu, web www.ams.ucsc.edu/~draper).
Table 1: **Distribution of number of hospitalizations in the IHGA study over a two-year period.**

<table>
<thead>
<tr>
<th>Group</th>
<th>Number of Hospitalizations</th>
<th>n</th>
<th>Mean</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>138 77 46 12 8 4 0 2</td>
<td>287</td>
<td>0.944</td>
<td>1.54</td>
</tr>
<tr>
<td>Treatment</td>
<td>147 83 37 13 3 1 1 0</td>
<td>285</td>
<td>0.768</td>
<td>1.02</td>
</tr>
</tbody>
</table>

for $i = 1, \ldots, n_C = 287$ and $j = 1, \ldots, n_T = 285$. But the last two columns of Table 1 reveal that the sample variance-to-mean ratios in the $C$ and $T$ 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

\[
\begin{align*}
(b_0^C, \sigma^2_C) & \sim \text{diffuse}, \\
(C_i | \lambda_i^C) & \sim \text{Poisson}(\lambda_i^C), \\
\log(\lambda_i^C) & = b_0^C + e_i^C, \\
(e_i^C | \sigma^2_e) & \sim N(0, \sigma^2_e),
\end{align*}
\]

and similarly for the treatment group. This model is more scientifically satisfying: each patient in the $C$ group has his/her own (latent) underlying rate of hospitalization $\lambda_i^C$, which may well differ from the underlying rates of the other $C$ patients because of unmeasured differences in factors such as health status at the beginning of the experiment.

Model (2), when extended in parallel to the $T$ 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 $\exp(b_0^T - b_0^C)$. However,

(a) nothing guarantees that the Gaussian mixing distribution in the last line of (2) 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 (2) 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 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. In this paper we contrast parametric and nonparametric models 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.

The plan of the paper is as follows. In Sections 2 and 3 we specify details of the parametric and Bayesian nonparametric models we study, respectively. Section 4 describes the simulation
setting, data sets and the results of comparing parametric and nonparametric models. It also includes detailed specifications of two more nonparametric models, a DP mixture model that introduces dependence in priors through a bivariate base distribution of DP, and another approach that models count data directly as draws from a Dirichlet process. Section 5 offers some extensions of the main models we examine and concludes with a brief discussion. In an Appendix we give a brief account of the Dirichlet process (DP) and Dirichlet process mixtures and provide computational and MCMC details to outline steps in the inference for obtaining posteriors of mixing distributions.

2 Parametric models

To fix notation, let \( Y_{1i} \) be the integer values in the sample under the control and \( Y_{2i} \) under the treatment, and denote by Poisson(\( \theta \)) the Poisson distribution (the distribution function or the probability mass function depending on the context) with mean \( \lambda = \exp(\theta) \), \( \theta \in R \).

Simple parametric modeling formulations include fixed-effects Poisson models,

\[
(Y_{ri}|\theta_r) \overset{\text{iid}}{\sim} \text{Poisson}(\theta_r)
\]

with independent priors \( G_{r0} \) for \( \theta_r, \ r = 1, 2 \), and random-effects Poisson models,

\[
(Y_{ri}|\theta_{ri}) \overset{\text{indep}}{\sim} \text{Poisson}(\theta_{ri})
\]

with independent priors \( G_{r0}, \ r = 1, 2 \). Under both (3) and (4), a standard choice for \( G_{r0} \) would be a normal distribution, \( N(\mu_r, \sigma_r^2) \), \( r = 1, 2 \). Both models are typically completed by adding normal and inverse-gamma hyperpriors on \( \mu_r \) and \( \sigma_r^2 \) respectively. We shall refer to (3) and (4) as models \( M_0 \) and \( M_1 \), respectively.

Posterior predictive inference under model (4) is straightforward; for a new data point \( y_0 \) we have \( [y_0|y] = \int[y_0|\theta_0]d[\theta_0|y] \). Furthermore, posteriors of mixing distributions, \( [G_r|\text{data}] \), are determined by the posterior samples of \( \mu_r \) and \( \sigma_r^2 \) allowing for a closed form for

\[
E(y|G_r) = \int \exp(\theta)dN(\theta; \mu_r, \sigma_r^2) = \exp(\mu_r + 0.5\sigma_r^2).
\]

The distribution of \( E(y|G_r) \) is obtained by computing \( E(y|G_r,b) = \exp(\mu_b + 0.5\sigma_b^2) \) for \( b = 1, B \). Obtaining distributions \([F(y|G)\mid\text{data}] \) and \([\text{median}(y)] \) proceeds exactly as described in the Appendix.

3 Bayesian nonparametric models

We consider two Bayesian nonparametric extensions to the Poisson random effects model. We treat random effects mixture distributions as unknown, and use Dirichlet process 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 (4) 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 distribution functions. We obtain the following DP mixture model:

\[
Y_{ri} | \theta_{ri} \quad \sim \quad \text{Poisson}(\theta_{ri}) \\
\theta_{ri} | G_r \quad \sim \quad G_r \\
G_r(\alpha_r, \mu_r, \sigma_r^2) \quad \sim \quad \text{DP}(\alpha_r, G_{r0}) \\
\alpha_r, \mu_r, \sigma_r^2 \quad \sim \quad [\alpha_r][\mu_r][\sigma_r^2],
\]

where the DP priors for \( G_r \), are independent and base distributions \( G_{r0} \) are \( N(\mu_r, \sigma_r^2) \), \( r = 1, 2 \). We shall refer to (5) as model \( M_2 \). The Pólya urn characterization of the DP (Blackwell and MacQueen, 1973) yields a useful marginalized version of (5) (integrating out \( G_r \), \( r = 1, 2 \) over their DP priors),

\[
Y_{ri} | \theta_{ri}, \ldots, \theta_{rn_r} \sim \text{Poisson}(\theta_{ri}) \\
\theta_{ri}, \ldots, \theta_{rn_r}, \alpha_r, \mu_r, \sigma_r^2 \sim [\alpha_r, \ldots, \alpha_r][\alpha_r, \mu_r, \sigma_r^2],
\]

where for \( r = 1, 2 \) the term \([\theta_{r1}, \ldots, \theta_{rn_r}, \alpha_r, \mu_r, \sigma_r^2]\) is equal to

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

and \( g_{r0} \) is the density of \( G_{r0} \). This expression specifies the prior probability model for the latent \( \theta_{r1}, \ldots, \theta_{rn_r} \) induced by the DP prior, and indicates that both parametric models (3) and (4) are limiting cases of the DP mixture model given in (5) and (6), 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 \( \theta_{ri} \).

This clustering in the prior of the \( \theta_{ri} \) results from the discreteness of the random distributions \( G_r \). 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 \( \Theta_r^* = (\theta_{r\ell}^* : \ell = 1, \ldots, n_{r*}) \) the vector of the distinct \( \theta_{ri} \). The vector of configuration indicators \( s_r = (s_{r1}, \ldots, s_{rn_r}) \), 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 \( \ell \), i.e., \( n_{r\ell} = |\{i : s_{ri} = \ell\}| \), \( \ell = 1, \ldots, n_{r*} \). MCMC algorithms that have been developed for DP mixture models (e.g., West, Müller and Escobar, 1994; Escobar and West, 1995; Bush and MacEachern, 1996; MacEachern and Müller, 1998; Neal, 2000) yield posterior samples from \([\theta_{r1}, \ldots, \theta_{rn_r}, \alpha_r, \mu_r, \sigma_r^2]\) and data, equivalently posterior samples from \([n_{r*}, s_r, \Theta_r^*, \alpha_r, \mu_r, \sigma_r^2]\) and data, where data = \( (Y_{r1}, \ldots, Y_{rn_r}) \).

The posterior predictive distribution for a future observable \( Y_{r\text{new}} \) under the control \( (r = 1) \) or the treatment \((r = 2)\) is given by

\[
[Y_{r\text{new}}|\text{data}] = \int \int \text{Poisson}(Y_{r\text{new}}|\theta_{r\text{new}}) d[\theta_{r\text{new}}] d[\eta_r]|d[\eta_r]|\text{data},
\]

(7)
where, based on the Pólya urn structure for the DP,

$$[\theta_r^{\text{new}} | \eta_r] = \frac{\alpha_r}{\alpha_r + n_r} G_{r0}(\theta_r^{\text{new}} | \mu_r, \sigma_r^2) + \frac{1}{\alpha_r + n_r} \sum_{\ell=1}^{n_r} n_r \delta_{\theta_r^{\text{new}}}(\theta_r^{\text{new}}),$$  

with $\eta_r$ collecting variables $n_r^*, s_r, \theta_r^*, \alpha_r, \mu_r, \sigma_r^2$. The integral expression in (7) indicates how to obtain a sample $(Y_r^{\text{new}} : b = 1, \ldots, B)$ from $[Y_r^{\text{new}} | \text{data}]$ using the posterior samples generated by MCMC. For each $b = 1, \ldots, B$ we first draw $\theta_r^{\text{new}}$ from (8) and then draw $Y_r^{\text{new}}$ from Poisson($Y_r^{\text{new}} | \theta_r^{\text{new}}$).

In order to obtain distributions $[G_r | \text{data}], [E(y) | G_r], [\text{median}(y) | G_r], r = 1, 2$, we follow exactly the steps given in the Appendix.

### 3.2 Stochastically ordered random-effects distributions

A special case of dependence for the random-effects distributions is induced by stochastic ordering, i.e., a prior assumption that $G_2(\theta) \leq G_1(\theta), \theta \in R$, (denoted by $G_1 \leqst G_2$), where $G_1$ and $G_2$ are the random effects distribution functions for the control and the treatment, respectively. (In our application, this probability order provides a formal way to capture the assumption that the number of hospitalizations under the treatment cannot be larger than under the control.) In certain applications, $G_1 \leqst G_2$ might be a natural assumption, e.g., if we know that the treatment yields improvement and we are interested in assessing the extent of improvement. Then, introducing the stochastic order restriction in the prior yields more accurate posterior predictive inference (e.g., narrower posterior interval estimates). In our study, a prior over the space $P = \{ (G_1, G_2) : G_1 \leqst G_2 \}$ yields one more nonparametric model that can be compared with (5) and (16).

A convenient way to build such a prior is by considering the subspace $P'$ of $P$ defined by

$$P' = \{ (G_1, G_2) : G_1 = H_1, G_2 = H_2 \},$$

where $H_1$ and $H_2$ are distribution functions on $R$, and placing independent DP priors on $H_1$ and $H_2$. Such a specification induces a nonparametric prior over $P'$, and hence over (a subset of) $P$, and allows for posterior inference based on extensions of standard MCMC methods for DP mixture models (Gelfand and Kottas, 2001; Kottas and Gelfand, 2001).

Note that a sample $\theta$ from $G_2 = H_1 H_2$ can be obtained by sampling independently $\theta_1$ from $H_1$ and $\theta_2$ from $H_2$, and then setting $\theta = \max(\theta_1, \theta_2)$.

In the two-sample case we assume $Y_{11}, \ldots, Y_{1m_1}$ i.i.d. from the mixture $F_1(\cdot; G_1) = \int k(\cdot; \theta_1) dG_1(\theta_1)$, and $Y_{21}, \ldots, Y_{2m_2}$ i.i.d. from $F_2(\cdot; G_2) = \int k(\cdot; \theta_2) dG_2(\theta)$. Letting $G_1 = H_1$ and $G_2 = H_1 H_2$ we have

$$F_2(\cdot; H_1 H_2) = \int \int k(\cdot; \max(\theta_1, \theta_2)) dH_1(\theta_1) dH_2(\theta_2),$$
and specifying independent DP priors on mixing distributions we obtain the following model

\[
\begin{align*}
Y_{1i}|\theta_i & \sim \text{Poisson}(\theta_i), i = 1, n_1 \\
Y_{2k}|\theta_1, n_1+k, \theta_2, k & \sim \text{Poisson}(\max(\theta_1, \theta_2)), k = 1, n_2 \\
\theta_1|H_1 & \sim \text{IID} H_1, i = 1, n_1 + n_2 \\
\theta_2|H_2 & \sim \text{IID} H_2, k = 1, n_2 \\
H_r|\alpha_r, \mu_r, \sigma_r^2 & \sim \text{DP}(\alpha_r H_0)
\end{align*}
\]

where the base distributions of Dirichlet processes, \(H_{10}\) and \(H_{20}\) are again normal with parametric priors on hyperparameters. We refer to (9) as model \(\mathcal{M}_3\).

The clustering of latent variables \((\theta_1, \ldots, \theta_{n_1+n_2})\) is again represented by the number of clusters \(n_r^*\), and a vector of distinct cluster values \(\theta_r^* = (\theta_{rj} : j = 1, n_r^*)\), and the indicator vector \(s = (s_1, \ldots, s_{n_1+n_2})\), with \(s_{1i} = \ell\) if and only if \(\theta_i = \theta_{\ell r}^*\), and \(n_\ell = |\{i : s_{1i} = \ell\}|\). The predictive distribution for \(\theta_{1\text{new}}\) is

\[
[\theta_{1\text{new}}|\eta_1] = \frac{\alpha_1}{\alpha_1 + n_1 + n_2} H_{10}(\theta_{1\text{new}}|\mu_1, \sigma_1^2) + \frac{1}{\alpha_1 + n_1 + n_2} \sum_{\ell=1}^{n_1} n_\ell \delta_{\ell r}(\theta_{1\text{new}}),
\]

and similarly, the predictive distribution for \(\theta_{2\text{new}}\) is

\[
[\theta_{2\text{new}}|\eta_2] = \frac{\alpha_2}{\alpha_2 + n_2} H_{20}(\theta_{2\text{new}}|\mu_2, \sigma_2^2) + \frac{1}{\alpha_2 + n_2} \sum_{\ell=1}^{n_2} n_\ell \delta_{\ell r}(\theta_{2\text{new}}).
\]

where \(\eta_r = (n_r^*, s_r, \theta_r^*, \alpha_r, \mu_r, \sigma_r^2)\). The posterior predictive distribution for a future \(Y_{1\text{new}}\) is given by

\[
[Y_{1\text{new}}|\text{data}] = \iint \text{Poisson}(Y_{1\text{new}}|\theta_{1\text{new}}) d[\theta_{1\text{new}}|\eta_1]d[\eta_1|\text{data}],
\]

and the posterior predictive distribution for \(Y_{2\text{new}}\) is

\[
[Y_{2\text{new}}|\text{data}] = \iiint \text{Poisson}(Y_{2\text{new}}|\max(\theta_{1\text{new}}, \theta_{2\text{new}})) d[\theta_{1\text{new}}|\eta_1]d[\theta_{2\text{new}}|\eta_2]d[\eta_1|\text{data}], d[\eta_2|\text{data}]
\]

To obtain posteriors of \(G_1\) and \(G_2\) we need first to compute posteriors for \(H_1\) and \(H_2\). In analogy to model \(\mathcal{M}_2\), we have \(H_{10}(\cdot|\psi_1) = \frac{\alpha_2}{\alpha_2 + n_1} H_{10}(\cdot|\psi_1) + \frac{1}{\alpha_2 + n_1} \sum_{i=1}^{n_1} \delta_{\ell r}(\cdot)\), and \(H_{20}(\cdot|\psi_2) = \frac{\alpha_2}{\alpha_2 + n_1 + n_2} H_{20}(\cdot|\psi_2) + \frac{1}{\alpha_2 + n_1 + n_2} \sum_{i=1}^{n_1 + n_2} \delta_{\ell r}(\cdot)\). We generate random draws from the posteriors \([H_r|\text{data}], r = 1, 2\), by following exactly the steps in the Appendix, and then we compute \([G_2|\text{data}]\) by multiplying pointwise realizations of \(H_1\) and \(H_2\) and set \(H_1 = G_1\). Computing all other posterior distributions of interest then proceeds in the same way as in the rest of the Appendix.

### 4 Simulation study

We have conducted a simulation study fitting the models \(\mathcal{M}_1\), \(\mathcal{M}_2\), and \(\mathcal{M}_3\) to four different data sets each with control \((D_{k,C})\) and treatment \((D_{k,T})\) parts, \(k = 1, 4\). In each case we obtained posterior predictive distributions along with the posteriors of random effects distributions.
4.1 Data sets

All data sets were of size $N = 300$ with data points drawn from Poisson($\exp(\theta_i)$), $i = 1, \ldots, N$, with the random-effects $\theta_i$-s generated as follows. For data set $D_1 = (D_{11}, Y_{21})$, we have $\theta_{11} \sim N(2.2, sd = 0.65)$, and $\theta_{21} \sim N(3.5, sd = 0.5)$. For data set $D_2$, $\theta_{1i}$-s are generated from a right skewed distribution (a four-component mixture of normals) and $\theta_{2i}$-s from a bimodal mixture of normals $0.5N(3.3, sd = 0.35) + N(5.8, sd = 0.42)$. The set $D_3$ is based on $\theta_{1i}$-s generated from $N(1.3, sd = 0.5)$ and $\theta_{2i}$-s from $N(2.2, sd = 0.5)$ distribution of $\theta_{1i}$-s is stochastically larger than the distribution of $\theta_{2i}$-s. Data set $D_4$ has $\theta_{1i}$-s drawn from $N(1.4, sd = 0.4)$ whereas, $\theta_{2i}$-s come from a bimodal mixture of two normals $0.5N(1.7, sd = 0.37) + 0.5N(3.3, sd = 0.52)$. Again, the distribution of random effects for the control group is stochastically larger than its treatment counterpart.

Figure 1 summarizes all of the data sets and the corresponding random effects.

4.2 Prior specification

Model $M_2$ reduces to the parametric random effects Poisson (PREP) model $M_1$ for $\alpha \rightarrow \infty$, what motivates specifying the same priors for hyperparameters. We have that $\theta = \ln(E(y))$ which we can regard as a “typical” $\theta$. Since $\theta \sim N(\mu, \sigma^2)$ and $\mu \sim N(\mu_0, \sigma_0^2)$ we set $\mu_0 = \ln(\bar{y})$ and set $\sigma_0$ to $K \times \ln(\max(y))$, where we take $K = 2$. Prior for $\sigma^2$ is an inverse gamma with infinite variance (shape parameter equal 2) and scale parameter $d$ equal to the mean of the distribution. Sensitivity analysis shows that values of $d$ around one result in stable inference and significant learning about $\sigma_2$. Similar values are also used for model $M_3$.

In DP mixture models the precision parameter $\alpha$ controls the distribution of latent variables $\theta_i$ and consequently the number of their distinct values (clusters), that is, the number of components in the mixture (details in Antoniak, 1974, and Escobar and West, 1995). We use a gamma prior with shape parameter 2 and such values for the scale parameter that allow for both small and large values for $\alpha$. For the size of data sets (300) used in this simulation study there is visible posterior learning for $\alpha$.

We follow the approach of Kottas and Gelfand (2001) when specifying the prior for model $M_3$.

4.3 Simulation results

We have provided full inference for the posterior predictive distribution of data and random mixing distributions $G$. Figure 2 shows an example of the posterior predictive distributions for the treatment part of data set $D_4$ and models $M_1$, $M_2$, and $M_3$. Prior predictive information is similarly diffuse for all three models, but clearly, only non-parametric models capture bimodality.

Perhaps more interesting is to see how the models can recover unknown mixing distributions $G$. CDF-s of the posteriors of mixing distributions, $[G|\text{data}]$ are shown for all data sets and three models in Figures 3–6. The parametric PREP model exhibits noticeably narrower 90% bands than both non-parametric models. Model $M_3$ in turn has somewhat narrower bands than $M_2$ since it incorporates a more informative prior on stochastic order. Ability to capture all the shapes of mixing distributions is apparent for non-parametric models, whereas the parametric
Figure 1: Simulation data sets, column i represents data set i. First and second rows give histograms of latent variables \( \theta \) used to generate integer samples, control and treatment, in third and fourth rows respectively. The fifth row shows empirical CDFs of latent variables, solid lines control, and dashed lines treatment.
Figure 2: Posterior and prior predictive data distributions for models $M_1$, $M_2$ and $M_3$ and $D_{4,T}$ dataset.
model fails for all data sets generated from skewed or bimodal random effects distributions. As an illustration of random distributions $G$ of random effects we show in Figure 7 samples from $DP$ of the corresponding CDF-s for all data sets.

Superior performance of non-parametric models on the data sets with random effects generated from a skewed and bimodal distributions (i.e. not from the parametric model) is confirmed by comparing corresponding values of log-score cross-validation criterion. Figure 8 shows larger values of both $LS_{FS}$ and $LS_{CV}$ for the non-parametric model on data sets $D_{2C}, D_{2T}$ and $D_{1T}$.

We have shown in the simulation study that the Bayesian nonparametric models clearly outperform their parametric counterparts in cases whenever the data are not typical samples from the parametric families. Even in the case when data are generated by the parametric models, the nonparametric models are as good.

5 Model extensions

We have considered other possible formulations of the BNP modeling approach in this (control–treatment) context. For the sake of completeness we outline two additional extensions in the following subsections.

5.1 Bayesian nonparametric fixed effects model

We can extend the simple formulation of fixed effects parametric model (3) by replacing the Poisson likelihood with an unknown and hence random distribution $F$ and specifying a Dirichlet stochastic process as its prior. Thus we obtain the following Bayesian nonparametric analogue of the parametric fixed effects model

$$ y_{ri}|F_r \overset{\text{ind}}{\sim} F(y_{ri}) $$
$$ F_r|\alpha_r, \theta_r \sim DP(\alpha_r, F_0) $$
$$ \alpha_r, \theta_r \sim [\alpha_r][\theta_r], $$

where $r = 1, 2, i = 1, \ldots, n$ and $F_0 = \text{Poisson}(\exp(\theta_r))$.

In the above model the random d.f. $F$ (we drop the subscript $r$ for clarity) is a mixture of DP-s ($F \sim \int DP(\alpha, F_0(\theta)) \, d[\alpha, \theta]$) and hence according to the Corollary (3.2') in Antoniak (1974) the posterior of $F$ is

$$ F|y \sim \int DP(\alpha', F_0'(\theta)) \, d[\alpha, \theta|y], $$

where $y = (y_1, \ldots, y_n), \alpha' = \alpha + n$, and $F_0'(t) = \frac{\alpha}{\alpha + n} F_0(t) + \frac{1}{\alpha + n} \sum_{i=1}^{n} 1[y_i, \infty)(t)$.

Lemma (1.1) in the same reference together with the discreteness of the base distribution $F_0$ allow us to derive the likelihood, a function of $\alpha$ and $\theta$, as in the following expression

$$ L(y; \alpha, \theta) = \frac{1}{\alpha^n(n)} \prod_{j=1}^{n^*} \alpha f_0(y_{j*}|\theta)[\alpha f_0(y_{j*}|\theta) + 1]^{(n_j-1)}, $$

where $n^*$ is the number of distinct values $y_{j*}$ in $y$, $n_j$ is size of the $j$-th cluster, $a^{(n)} = a(a + 1) \ldots (a + n - 1)$ and $f_0(x|\theta) = (x!)^{-1} \exp(\theta x - e^\theta)$.
Figure 3: Posterior MCMC estimates of the random mixing distributions for models $\mathcal{M}_1$ and $\mathcal{M}_2$ (first and second column respectively), and data sets $D_{1,C}$ and $D_{1,T}$ (first and second row respectively).
Figure 4: Posterior MCMC estimates of the random mixing distributions for models $M_1$ and $M_2$ (first and second column respectively), and data sets $D_{2C}$ and $D_{2T}$ (first and second row respectively).
Figure 5: Posterior MCMC estimates of the random mixing distributions for models $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$ (first, second, and third rows respectively), and data sets $D_{3,C}$ and $D_{3,T}$ (first and second columns respectively).
Figure 6: Posterior MCMC estimates of the random mixing distributions for models $M_1$, $M_2$ and $M_3$ (first, second, and third row respectively), and data sets $D_{AC}$ and $D_{AT}$ (first and second columns respectively).
Figure 7: Random samples from (G|data) for model \(W^2\) and four datasets (G first column, J second column).
Figure 8: $L_{SCV}$ (left panel) vs $L_{FS}$ (right panel) for models $M_1$ and $M_2$ and all data sets, $D_{1,C}$, $D_{1,T}$, $\ldots$, $D_{4,C}$, $D_{4,T}$.
The joint posterior for $\alpha$ and $\theta$ can be written as

$$[\alpha, \theta | y] \propto \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)} \prod_{j=1}^{n'} \frac{\Gamma(\alpha f_0(y_j^* | \theta) + n_j)}{\Gamma(\alpha f_0(y_j^* | \theta))} [\alpha | \theta],$$

where priors for $\alpha$ and $\theta$ are Gamma and normal distributions respectively. To obtain samples from the posterior $[\alpha, \theta | y]$ we use a symmetric random walk on $(\log(\alpha), \theta)$.

Furthermore, the posterior predictive distribution is

$$[y^{\text{new}} | y] = \int [y^{\text{new}} | F] d[F | y]$$

where $[F | y]$ is a mixture of DP-s given in (11).

The expressions (14) and (11) indicate how to sample from the posterior predictive distribution of the data. For each $l = 1, L$, where $L$ is the size of the posterior sample,

1. obtain an instance of $F$ (which is a CDF) from $[F | y]$ by sampling from $\text{DP}(\alpha^{(l)} + n, F_0^{(l)}(\theta^{(l)}))$;  

2. given $F$ from the step 1. generate $y^{\text{new}}$ using inverse CDF method.

Alternatively, since $\int p(Y = y^{\text{new}} | F) d[F | y] = E[F | y][p(Y = y^{\text{new}} | F)]$, instead of sampling $y^{\text{new}}$ in the second step above, we can estimate the posterior predictive probability at any point $y_0$ with

$$p(Y = y_0 | y) \approx \frac{1}{L} \sum_{l=1}^{L} F^{(l)}(y_0) - F^{(l)}(y_0 - 1),$$

where $F^{(l)}$ is the sample obtained from $\text{DP}(\alpha^{(l)} + n, F_0^{(l)}(\theta^{(l)}))$, for $y_0 > 0$. At $y_0 = 0$ (15) reduces to $p(Y = 0 | y) \approx \frac{1}{L} \sum_{l=1}^{L} F^{(l)}(0)$.

### 5.2 Bivariate prior probability model for the random-effects distributions

A more general nonparametric model than $\mathcal{M}_2$ can be developed by introducing dependence in the prior probability model for the random-effects distributions, i.e., by placing a nonparametric prior over the space of bivariate distribution functions for the pair of random effects $(\theta_{1i}, \theta_{2i})$.

We want the model to allow for dependence in the prior for the random-effects distributions. In fact, its formulation is in the spirit of ANOVA dependent DP modeling (De Iorio, et al., 2004), although it would be somewhat different if we were to follow exactly the approach in the above reference. Specifically, denoting by $Y_i$, $i = 1, \ldots, n_1 + n_2$, the counts in the entire data set, an alternative to model $\mathcal{M}_2$ would be

$$\begin{align*}
Y_i | \varphi_i & \overset{\text{ind.}}{\sim} \text{Poisson}(\varphi_i^T d_i), \ i = 1, \ldots, n_1 + n_2 \\
\varphi_i | F & \overset{i.i.d.}{\sim} F, \ i = 1, \ldots, n_1 + n_2 \\
F | \alpha, \psi & \sim \text{DP}(\alpha F_0(\psi)) \\
\alpha, \psi & \sim [\alpha | \psi],
\end{align*}$$

(16)
where \( \varphi_i = (\varphi_{0i}, \varphi_{1i}), d_i = (1, 0) \) for the control, \( d_i = (1, 1) \) for the treatment, and \( F_0 \) can be defined by two independent normals.

Nonparametric model (16), to which we refer as model \( M_4 \), relates directly with standard parametric random effects models. In fact, it is a generalization of the PREP model and is a natural choice if several categorical covariates are available. The model \( M_4 \) places no restrictions on the number of covariates and it does not require them to be continuous.

A marginalized version of (16) results after integrating out \( G \) over its DP prior:

\[
(Y_i | \varphi_i) \overset{\text{ind}}{\sim} \text{Poisson}(\varphi_i^T d_i), i = 1, \ldots, n_1 + n_2 \\
(\varphi_1, \ldots, \varphi_{n_1+n_2} | \alpha, \psi) \overset{\text{iid}}{\sim} [\varphi_1, \ldots, \varphi_{n_1+n_2} | \alpha, \psi] \\
\overset{\text{iid}}{\sim} [\alpha | \psi],
\]

(17)

Again, the discreteness of the DP prior induces a number, \( n^{*} \), of distinct elements \( (\varphi_{0*}, \varphi_{1*}) = ((\varphi_{0\ell}, \varphi_{1\ell}) : \ell = 1, \ldots, n^{*}) \), in the vector \( ((\varphi_{01}, \varphi_{11}), \ldots, (\varphi_{0n}, \varphi_{1n})) \), which therefore can equivalently be represented by \( n^{*}, (\varphi_{0*}, \varphi_{1*}), s = (s_1, \ldots, s_n) \), with \( s_i = \ell \) if and only if \( (\varphi_{0i}, \varphi_{1i}) = (\varphi_{0\ell}, \varphi_{1\ell}) \), and \( n_\ell = |\{i : s_i = \ell\}| \). Letting \( \eta = (n^{*}, s, (\varphi_{0*}, \varphi_{1*}), \alpha, \phi) \), the predictive distribution for \( \varphi^{\text{new}} = (\varphi_{0\text{new}}, \varphi_{1\text{new}}) \), associated with future observation \( Y^{\text{new}} \), is given by

\[
[(\varphi^{\text{new}}_{0}, \varphi^{\text{new}}_{1}) | \eta] = \frac{\alpha}{\alpha + n} G_0(\varphi^{\text{new}}_{0}, \varphi^{\text{new}}_{1} | \phi) + \frac{1}{\alpha + n} \sum_{\ell=1}^{n^{*}} n_\ell \delta_{(\varphi_{0\ell}, \varphi_{1\ell})}(\varphi^{\text{new}}_{0}, \varphi^{\text{new}}_{1})
\]

and the posterior predictive distribution for \( Y^{\text{new}} \) associated with \( d_i \) is

\[
[Y^{\text{new}} | \text{data}] = \int \int \text{Poisson}(\varphi^{\text{new}}_{0} + \varphi^{\text{new}}_{1} d_{1i})[(\varphi^{\text{new}}_{1}, \varphi^{\text{new}}_{2}) | \eta][\eta | \text{data}].
\]

In this document we do not provide comparisons with this model as it requires somewhat different setting as a consequence of the bivariate random effects distribution. However, it merits the mention and further analysis as a generalization of the model (5), capable of handling dependence in the random effects distributions.

### 5.3 Discussion

[to be supplied]

### Appendix

In order to facilitate the presentation of our Bayesian nonparametric approach to modeling count data, here we review basic definitions and results on Dirichlet processes and Dirichlet process mixtures. The theory of Dirichlet processes was established by Ferguson (1973, 1974), Blackwell (1973), and Blackwell and MacQueen (1973), and extended by, among others, Cifarelli and Regazzini (1990) and Lo (1991), with review papers by (e.g.) Walker et al. (1999) and Mueller and Quintana (2004) giving additional references.

**The Dirichlet process.** A Dirichlet process (DP) is a stochastic process with sample paths which are cumulative distribution functions (CDFs). Let \( \Omega \) be a sample space and \( \mathcal{F} \) a \( \sigma \)-field of subsets of
Parametric and nonparametric Bayesian model specification

Following the work of Ferguson (1973, 1974) we define a Dirichlet process as a stochastic process $Q = \{Q(\omega, A) : \omega \in \Omega, A \in \mathcal{F}\}$ with sample paths $\{Q_\omega(A) = Q(\omega, A), \forall A \in \mathcal{F}\}$ that are probability measures on $(\Omega, \mathcal{F})$, such that for any finite measurable partition $(A_1, \ldots, A_n)$ of $\Omega$ the random vector $(Q(A_1), \ldots, Q(A_n))$ has a Dirichlet distribution with parameters $[\alpha Q_0(A_1), \ldots, \alpha Q_0(A_n)]$, $\alpha \in \mathbb{R}$. Here $Q_0(A)$, a constant, and $Q(A)$, a random variable, denote the probability of event $A$ under $Q_0$ and $Q$ respectively. $Q$ is also called 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 $V[Q(A)] = Q_0(A)(1 - Q_0(A))/(\alpha + 1)$.

The measure $Q_0$ is considered the center of the process and $\alpha$ is interpreted as a precision parameter. A DP is defined on the space of probability measures but we often use the term distribution instead as the two are equivalent, albeit in different probability spaces. For example, when $\Omega = \mathbb{R}$ and $A = (-\infty, x), x \in \mathbb{R}$, then $Q(A) = G(x)$ has a Beta distribution with parameters $\alpha G_0(x)$ and $\alpha[1 - G_0(x)]$ and hence $E[G(x)] = G_0$ and $V[G(x)] = G_0(x)(1 - G_0(x))/(\alpha + 1)$, where $G$ and $G_0$ are distributions on $\mathbb{R}$. For larger precision parameters $\alpha$ a realization $G$ from a DP is expected to be closer to the base distribution $G_0$ than for smaller values of $\alpha$.

It turns out that random realizations (sample paths) from Dirichlet process are almost surely (a.s.) discrete distributions, and Sethuraman and Tiwari (1982) provide a constructive definition of DPs as a countable mixture of point masses. Specifically, let $\{z_k, k = 1, 2, \ldots\}$ and $\{\psi_j, j = 1, 2, \ldots\}$ be independent sequences of IID random variables with $z_k \sim \text{Beta}(1, \alpha)$ and $\psi_j \sim G_0$. Defining $w_1 = z_1, w_i = z_i \prod_{k=1}^{i-1}(1 - z_k), i = 2, 3, \ldots$ makes a realization $G$ from $\text{DP}(\alpha, G_0)$ a.s. of the form

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

where $\delta_x(\cdot)$ denotes a point mass at $x$. We say that the DP generates, with probability one, distributions that can be represented as countable mixtures of point masses, with locations $\psi_i$ drawn independently from $G_0$ and weights $w_k$ generated according to a “stick breaking” procedure based on IID draws $z_k$ from Beta$(1, \alpha)$.

An attractive property of the DP, when used as a prior on a space of distribution functions, is that the posterior may be obtained through straightforward conjugate updating. Ferguson (1973) proved that for an IID sample $y = (y_1, \ldots, y_n)$ from $G$ and with a DP prior on $G$, i.e., $G \sim \text{DP}(\alpha, G_0)$, the posterior distribution of $(G|y)$ is $\text{DP}(\alpha + n, G_0)$, where

$$\tilde{G}_0(\cdot) = \frac{\alpha}{\alpha + n} G_0(\cdot) + \frac{1}{\alpha + n} \sum_{i=1}^{n} 1_{[y_i, \infty)}(\cdot).$$

**Dirichlet process mixtures.** Almost sure discreteness of realizations from a Dirichlet process imposes modeling limits, so if we need to model continuous data, we are naturally lead to consider some kind of mixture models. In Dirichlet process mixtures (DPM) we avoid modeling the data as a direct draw from a random distribution. Instead, data points $y_i$ are regarded as independent draws from a parametric kernel $k$, conditional on latent variables $\theta_i$, which are in turn given as IID draws from a random distribution $G$, as in the following DPM (hierarchical) model:

$$\begin{align*}
(y_i|\theta_i) & \overset{\text{indep}}{\sim} k(\cdot; \theta_i) \\
(\theta_i|G) & \overset{\text{IID}}{\sim} G \\
G & \sim \text{DP}(\alpha, G_0).
\end{align*}$$


The equivalent formulation of the DP mixture model below, in which the latent $\theta_i$ have been integrated out, emphasizes the role of the random CDF $G$ as the mixing distribution:

$$
(y_i | G) \overset{\text{ID}}{\sim} F(\cdot; G) = \int k(\cdot; \theta) \, dG(\theta)
$$

$$
G \sim \text{DP}(\alpha, G_0),
$$

where $i = 1, \ldots, n$, $k(\cdot; \theta)$ is a parametric kernel and the base distribution $G_0 = G_0(\cdot; \psi)$ is allowed to be indexed by a vector of parameters $\psi$.

The main theoretical results about Dirichlet process mixtures can be found in the work of Antoniak (1974), followed by (e.g.) Lo (1984), Kuo (1986), and Lavine and Mockus (1995). MCMC-based posterior inference in DP mixture models was made practical in the work of Escobar and West (1995), and was further extended by (e.g.) Bush and McEachern (1996) and Neal (2000).

**Posterior inference for random mixing distributions.** For the DP mixture model (21) we are interested in obtaining posterior inference for the mixing distribution $G$, the mixture distribution $F$ and their associated functionals. Following Ferguson (1973) and Antoniak (1974), we have that $[\theta, \alpha, \psi, G | \text{data}] \propto [G | \theta, \alpha, \psi] [\theta, \alpha, \psi | \text{data}]$, using the bracket notation of Gelfand and Smith (1990) to denote densities or distributions. Therefore

$$
[G | \text{data}] = \int [G | \theta, \alpha, \psi] \, d[\theta, \alpha, \psi | \text{data}],
$$

meaning that the posterior distribution of the random $G$ (where $G = \text{DP}$) is a mixture of DPs $([G | \theta, \alpha, \psi])$.

Here $[G | \theta, \alpha, \psi]$ is a DP with updated parameters $\alpha' = \alpha + n$ and $G_0'(\cdot | \psi) = \frac{\alpha}{\alpha + n} G_0(\cdot | \psi) + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{\theta_i} (\cdot)$, where $\theta = (\theta_1, \ldots, \theta_n)$ and $\psi$ collects parameters of the base distribution. Then, for any $t \in \mathbb{R}$,

$$
E[G(t) | \text{data}] = \int E([G(t) | \theta, \alpha, \psi]) \, d[\theta, \alpha, \psi | \text{data}],
$$

yielding $E[G(t) | \text{data}] = \int \tilde{G}_0(t) \, d[\theta, \alpha, \psi | \text{data}]$, where

$$
\tilde{G}_0(t; \theta, \alpha, \psi) = \frac{\alpha}{\alpha + n} G_0(t; \psi) + \frac{1}{\alpha + n} \sum_{i=1}^{n} 1_{[\theta_i, \infty)} (t).
$$

Now, given a posterior sample $\{\theta^*_k, \alpha^*_k, \psi^*_k, b = 1, \ldots, B\}$, we can directly obtain a Monte Carlo estimate of $E[G(t) | \text{data}]$ for a fixed $t$ from (23) and (24).

Going beyond estimating only the posterior expectation of a random distribution, we want to obtain its entire posterior, $[G | \text{data}]$. To that end we need random samples $[G(t_1), \ldots, G(t_L)]$ from $[G | \theta, \alpha, \psi]$ in (22) which can be drawn from an ordered Dirichlet distribution with parameters determined by $\theta, \alpha$, and $\psi$. Equivalently, we sample $(u_1, \ldots, u_L)$ from a Dirichlet distribution with parameters $\bar{\alpha} G_0(t_1), \bar{\alpha}(G_0(t_2) - G_0(t_1)), \ldots, \bar{\alpha}(1 - G_0(t_L)), \text{where } \bar{\alpha} = \alpha + n$, and $G_0$ is given in (24). Now, setting $[G(t_1), G(t_2), \ldots, G(t_L)] = (u_1, u_1 + u_2, \ldots, \sum_{j=1}^{L} u_j)$, we obtain a draw from $[G | \text{data}]$.

Repeating the above steps for $B$ available posterior parameter samples we obtain the entire posterior distribution of the random mixing distribution $G$, from which we may compute any associated quantities of interest such as (e.g.) the mean, median, and 5th and 95th percentiles. This approach is in contrast with Gelfand and Kottas (2003), who use Sethuraman’s construction of the DP to draw from $[G | \theta, \alpha, \psi]$. 


Once we have available samples from $[G|\text{data}]$ we can compute estimates of $E(y|G)$, and in fact, obtain the entire distribution $[E(y|G)|\text{data}]$. For models $\mathcal{M}_1$ and $\mathcal{M}_2$ we have $E(y|G) = E[\theta|G]E(y|\theta) = \int \exp(\theta) \, dG(\theta)$, which is estimated by
\[
    u_b = \sum_{l=1}^{L} \exp(t_l)[G_b(t_l) - G_b(t_l^-)],
\]
for a fixed MCMC iteration $b$. An estimate of the distribution $[E(y|G)|\text{data}]$ is obtained by computing $u_b$s for all MCMC posterior samples, $b = 1, \ldots, B$.

Before obtaining the posterior distribution of the median of $(y|G)$ we need to make one more step and compute the posterior distribution of the CDF-at-a-point functional $[F(y; G)|\text{data}]$ (Gelfand and Mukhopadhyay, 1995). For a specific fixed point $y_0$ and a specific iteration $b$ we have $F(y_0; G_b)|\text{data} = \int K(y_0; \theta) \, dG_b(\theta)$ at a fixed point $y_0$, which can be estimated by the following sum on a grid of values $t_1, \ldots, t_L$ for $\theta$. Therefore, $F(y_0; G_b)|\text{data} = \sum_{l=1}^{L} K(y_0; t_l)[G_b(t_l) - G_b(t_l^-)]$. Repeating the previous calculation for a grid of data points $y_1 < y_2 < \ldots < y_K$ and for samples $b = 1, \ldots, B$ we now obtain a $B \times K$ matrix of realizations $F(y_k; G_b)$. Each row $F(y_1; G_b), \ldots, F(y_K; G_b)$ for a fixed $b$ is a random realization of the cdf $F(\cdot|G)$ from $[F(\cdot; G)|\text{data}]$, and each column is a sample from the posterior $[F(y_k; G)|\text{data}]$. We obtain the median of $(y|G)$ by inverting the row vectors.

References


