A Nonparametric Bayesian Modeling Approach for Cytogenetic Dosimetry

Athanasios Kottas

Institute of Statistics and Decision Sciences, Duke University

Durham, NC 27708-0251, USA

email: thanos@stat.duke.edu

Márcia D. Branco

Departamento de Estatística, IME, Universidade de São Paulo

Cxa Postal 66281, 05315-970, São Paulo, SP, Brasil

email: mbranco@ime.usp.br

and

Alan E. Gelfand

Department of Statistics, University of Connecticut

Storrs, CT 06269-4120, USA

email: alan@stat.uconn.edu

KEY WORDS: Auxiliary variables; Dirichlet process; Dose-response; Logistic regression; Polytomous response.

SUMMARY. In cytogenetic dosimetry, samples of cell cultures are exposed to a range of doses of a given agent. In each sample, at each dose level, some measure of cell disability is recorded. The objective is to develop models which explain cell response to dose. Such models can be used to predict response at unobserved doses. More importantly, such models can provide inference for unknown exposure doses given the observed responses. Typically, cell disability is viewed as a Poisson count but in the present work a more appropriate response is a categorical classification. In the literature, modeling in this case is very limited. What exists is purely parametric. We propose a fully Bayesian nonparametric approach to this problem. We offer comparison with a parametric model through a simulation study and the analysis of a real dataset modeling blood cultures exposed to radiation where classification is with regard to number of micronuclei per cell.

1 Introduction

Cytogenetic dosimetry is the particular area of dose-response modeling which is concerned with the relationship between dose as some form of exposure to radiation and response as some measure of genetic aberration. Such relationships are used to address two questions of primary interest, (i) prediction of response at unobserved doses/exposure levels and, more importantly, (ii) inference for unknown exposures given observed responses. The latter inversion problem distinguishes cytogenetic dosimetry from usual dose-response settings in that, while response is usually accurately observed, exposure is typically very difficult to measure.

Cytogenetic dosimetry has been studied both *in vivo* and *in vitro*. The former usually involves human exposures, with the response being chromosomal aberration. Typically, uncomfortable extrapolation arises. Measurements are at lower doses while interest is in the relationship at higher levels (Bender et al., 1988). In the in vitro setting, samples of cell cultures of human lymphocytes are exposed to a range of doses of a given agent. In each sample, at each dose level, again, some measure of chromosomal aberration or cell disability is recorded.

Often it is a count which is customarily assumed to follow a Poisson distribution. Thus, Poisson regressions on dose are standard with a so-called *linear-quadratic* model, $\lambda(X) = \alpha X + \beta X^2$ where λ is the Poisson intensity and X the dose, being predominant (Frome and DuFrain, 1986). The inversion or statistical calibration problem is straightforward in this case (Osborne, 1991).

In some cases a more appropriate response is a categorical classification. The case of binary response yields the familiar bioassay problem. In the cytogenetic dosimetry literature, the polytomous response case has been examined, as in, for example, Madruga, Pereira and Rabello-Gay (1994) and Madruga et al. (1996) but exclusively under parametric models.

We propose a flexible nonparametric modeling approach to this problem and demonstrate how it can be used to address the questions of interest. We adopt a Bayesian perspective in our framework which attractively provides an entire posterior distribution for all predictions. Gelfand and Kuo (1991, p. 662-664) present some initial work in this regard. The Bayesian perspective is particularly helpful in making the inversion problem under categorical response well defined. That is, simultaneous inversion of a set of response curves, one for each classification, to obtain a common dose, does not provide a uniquely defined solution. However, modeling the distribution of the unknown risk as a function of exposure and then the categorical responses given this risk, directly provides a posterior for the unknown exposure given the observed response vector and all of the other data. Mukhopadhyay (2000) offers a nonparametric Bayesian treatment of the inversion problem in the binary response case where only one response curve is involved.

Suppose we assume a multinomial model for the categorical response with a conjugate Dirichlet distribution on the probabilities. Following Aitchison and Shen (1980), the log ratio transformed probabilities follow a multivariate normal distribution, leading to a posterior which is multivariate normal on the transformed scale (Pereira and Pericchi, 1990). Such modeling can not handle prediction or inversion. In order to do so, Madruga et al. (1994) introduce a parametric model for the log ratio transformed probabilities, as a function of dose.

In 1986, a scientific committee was established, at the request of the National Cancer Institute, to assess the status of cytogenetic procedures to detect and quantify previous exposures to radiation. Bender et al. (1988) summarizes this committee's findings recording its adamantly Bayesian stance with various assertions on p. 139-140 such as, "the committee has used the Bayesian approach to dose estimation simply because it is the only approach which completely answers the questions with which we are faced". With the current ability to investigate rich nonparametric models, the committee's stance seems even more appropriate.

In section 2 using a data set taken from Madruga et al. (1996), we present a parametric Bayesian analysis using a simple logistic structure to model probabilities. We use this for comparison with the subsequent nonparametric analysis. Our application features ordinal classifications suggesting natural cumulation of probabilities. In section 3 we present a fully nonparametric approach, which accommodates this order. Section 4 provides a simulation study to examine the performance of the nonparametric model, in particular, compared to the parametric model. Finally, in section 5 we offer further comparison, summary and related discussion.

2 The Data and a Parametric Model

The data we study is a portion of a larger set where blood samples from individuals were exposed in vitro to ^{60}Co radiation with doses of 20, 50, 100, 200, 300, 400 and 500 cGy (centogram). Lymphocyte cultures were prepared for a cytokinesis-block micronucleus assay and analyzed for the presence of mono- and binucleated cells with none, one, and two or more micronuclei (MN). The use of these three classifications rather than the actual counts arises because, when there are multiple micronuclei, the assayers find it difficult to count the exact number. More details and the full data set are provided in Madruga et al. (1996). Here we confine ourselves to the binucleated cells from the two healthy older subjects.

(Table 1 here)

With categorical response at each dose level, a multinomial model is the customary assumption. That is, for dose levels d_i, $i = 1, ..., k$ and classifications $j = 1, ..., r$, we assume $\mathbf{Y}_i = (Y_{i1}, ..., Y_{ir})$, the vector of observed counts in each response class, is $Mult(n_i, \mathbf{p}_i)$ where n_i is the number of cells studied at the i^{th} dose level and $\mathbf{p}_i = (p_{i1}, ..., p_{ir})$ denotes the unknown probabilities of each classification at each dose. The data we used are given in Table 1 where $k = 7$ and $r = 3$. Also given are the sample estimates of at least two micronuclei, i.e., $\widehat{\eta}_{i1} \equiv y_{i1}/(y_{i1} + y_{i2} + y_{i3})$ and at least one micronuclei, i.e., $\widehat{\eta}_{i2} \equiv (y_{i1} + y_{i2})/(y_{i1} + y_{i2} + y_{i3})$.

The inferential objectives are to learn about the p_{ij}. In addition, at a new dose level d_0, we would like to estimate the vector $\mathbf{p}_0$ and, hence, for a specified n_0 be able to predict $\mathbf{Y}_0$. Inversely, we would like to estimate an unobserved dose level given an observed $\mathbf{Y}_0$ (hence n_0). The latter two problems presume continuity of $\mathbf{p}$ in dose d.

Parametric models specify $p_{ij} = g(d_i; \theta_j)$ for a specified function g with θ_j being the parameters associated with classification j and, of course, $\sum_j p_{ij} = 1$. We illustrate with a very simple version employing the logit, setting

$$\log \frac{p_{ij}}{p_{ir}} = \beta_{1j} + \beta_{2j} \log d_i. \tag{1}$$

Nonlinear forms for the logit are discussed in Madruga et al. (1996). In fact, these authors work with the larger dataset and introduce an effect for whether the cells are mononucleated or binucleated.

Bayesian inference under (1) is discussed, for example, in Johnson and Albert (1999). Model fitting is routine using the BUGS software (Spiegelhalter et al., 1995). With a flat prior on the (β_{1j}, β_{2j}), we have no hierarchical structure. It is straightforward to demonstrate that a proper posterior results if, for each j, there are at least two Y_{ij}'s such $0 < Y_{ij} < n_i$. See Gelfand and Sahu (1999) and further references therein. More generally, a bivariate normal prior could be introduced for (β_{1j}, β_{2j}). If appropriate, the (β_{1j}, β_{2j}) might be assumed i.i.d., adding hyperparameters and a hyperprior, creating a hierarchical model. Conditions for posterior propriety are given in Hobert and Casella (1996).

For the data in Table 1 we adopt the flat prior assumption, presenting posterior summaries for the β_{1j} and β_{2j} in Table 2. Here, $j = 1$ denotes the event "two or more MN", $j = 2$, "exactly one MN", $j = 3$, "no MN". Encouragingly, both β_{21} and β_{22} are significantly positive; the chance of cell aberration increases in dose. The posteriors for the resultant p_{ij} are summarized in Table 3. Prediction at a new dose d_0 is straightforward, merely requiring the posteriors for $g(d_0; \theta_j)$. In fact, of most interest is prediction of the probability of two or more MN at d_0, i.e., $g(d_0, \theta_1)$ and the probability of at least one MN at d_0, i.e., $g(d_0, \theta_1) + g(d_0, \theta_2)$. The means of these predictive posteriors are plotted as a function of d_0 in Figure 2. Interval estimates are directly available from the posterior samples but are not shown. They are in accord with those for the p_{ij} in Table 3.

(Table 2 here)

Madruga et al. (1996) consider the inversion problem for a healthy older subject presenting the count vector $\mathbf{Y}_0 = (316, 801, 1310)$ at an unknown d_0. (Inspection of Table 1 suggests a d_0 larger than 500, yielding an extrapolation problem.) In particular, they obtained a simultaneous 95% credibility set for the category probabilities associated with $\mathbf{Y}_0$ and did an approximate inversion under their nonlinear model to obtain a credible interval for the unknown dose. They obtained $(815, 1210)$ as a roughly 95% credible interval for d_0. We consider inversion for $\mathbf{Y}_0$ under the model in (1). We also attempt to validate our inversion taking $\mathbf{Y}_0 = (32, 114, 939)$, in fact, the observed data at $d = 100$. In each case all that is required is to add another unknown to the model, d_0, with another term in the product which completes the likelihood. For illustration, we adopt a prior on $\log d_0$ which is $N(4.96, 9)$, i.e., it is centered at the average log dose for our sample with very large variance. It produces a 6σ- range on the log scale of $(-4.04, 13.96)$, hence a range $(0.02, 1.15 \times 10^6)$ on the dose scale. We obtain the point (posterior median) and 95% equal tail interval estimate for d_0 provided in Table 4. Due to the large number of blood cells observed at d_0, the data essentially overwhelms this prior. However, the model in (1) may not be appropriate. Indeed, our interval estimate differs considerably from the nonequal tail interval estimate obtained under the nonlinear model of Madruga et al. (1996).

It is noteworthy that, with the assumed flat prior on the β_{1j} and β_{2j}, if in addition, we take a flat prior on an unbounded range for $\log d_0$, an improper posterior results. This is easily seen, for instance, in the case $r = 2$. Then, the posterior for β_1, β_2 and $\log d_0$ is proportional to $\left\{\prod_{i=1}^{k} e^{(\beta_1+\beta_2 \log d_i) y_i} / (1 + e^{\beta_1+\beta_2 \log d_i})^{n_i}\right\} e^{(\beta_1+\beta_2 \log d_0) y_0}$

$/(1+e^{\beta_1+\beta_2 \log d_0})^{n_0}$. Reparametrizing to β_1, β_2 and $z_0 = \beta_2 \log d_0$ and then integrating over z_0 yields the form $c(\beta_1)\beta_2^{-1} \prod_{i=1}^{k} e^{(\beta_1+\beta_2 \log d_i)y_i}/(1+e^{\beta_1+\beta_2 \log d_i})^{n_i}$ which is not integrable with respect to β_2.

3 A Fully Nonparametric Approach

Now, we propose a fully nonparametric model which is also straightforward to fit. With $\eta_{ij} = \sum_{\ell=1}^{j} p_{i\ell}$, $j = 1, ..., r-1$, we define

$$\eta_{ij} = F_j(\log d_i) = \prod_{\ell=1}^{r-j} G_\ell(\log d_i). \tag{2}$$

In (2), each G_ℓ is a c.d.f. so that η_{ij} is increasing in d_i and $\eta_{ij} < \eta_{i,j+1}$. Evidently, $p_{i1} = \eta_{i1} = \prod_{\ell=1}^{r-1} G_\ell(\log d_i)$ and $p_{ir} = 1 - \eta_{i,r-1} = 1 - G_1(\log d_i)$. But also for $j = 2,...,r-1$, $p_{ij} = \eta_{ij} - \eta_{i,j-1}$ and thus $p_{ij} = (1 - G_{r-j+1}(\log d_i)) \prod_{\ell=1}^{r-j} G_\ell(\log d_i)$.

The G_ℓ's are arbitrary and unknown and hence, under a Bayesian framework, taken to be random. They are introduced because we propose to model them as independent random functions. Describing the F_j's through products of G_ℓ's is an assumption which is made for convenience. It is easier to model an independent set of unknown random distributions than to model the F_j directly. To simplify notation, we let $G_\ell(\log d_i) = q_{\ell i}$.

A nonparametric approach for modeling the G_ℓ is to assume that, a priori, each arises at random from a distribution placed on a rich set of distributions. (This differs from a parametric approach where each G_ℓ would be assumed from some standard parametric family of distributions.) A flexible and computationally convenient prior on the class of distributions is to model each G_ℓ as a realization from a Dirichlet process (Ferguson, 1973, 1974), i.e., $G_\ell \sim DP(\alpha_\ell G_{0\ell})$ where $G_{0\ell}$ is a known distribution and $\alpha_\ell > 0$ is a given precision parameter. This process is defined such that, for any partition of R^1, say $(B_1, B_2, ..., B_t)$, the random vector of probabilities

$$(G_\ell(B_1), ..., G_\ell(B_t)) \sim Dir(\alpha_\ell G_{0\ell}(B_1), ..., \alpha_\ell G_{0\ell}(B_t)).$$

The interpretation is essentially that $EG_\ell = G_{0\ell}$ and that G_ℓ tends to be closer to $G_{0\ell}$ as α_ℓ increases.

Assuming the doses are ordered, i.e., $d_1 < d_2 < \cdots < d_k$ and letting I_i be the interval $(\log d_{i-1}, \log d_i]$, $i = 2, ..., k$ with $I_1 = (-\infty, \log d_1]$ and $I_{k+1} = (\log d_k, \infty)$ we immediately induce a distribution on $\mathbf{q}_\ell = (q_{\ell 1}, ..., q_{\ell k})$, i.e.,

$$(q_{\ell 1}, q_{\ell 2} - q_{\ell 1}, ..., q_{\ell k} - q_{\ell,k-1}, 1 - q_{\ell k}) \sim Dir(\alpha_\ell q_{0\ell 1}, \alpha_\ell(q_{0\ell 2} - q_{0\ell 1}), ..., \alpha_\ell(1 - q_{0\ell k})), \tag{3}$$

where $q_{0\ell i} = G_{0\ell}(\log d_i)$ and thus $q_{0\ell i} - q_{0\ell,i-1} = G_{0\ell}(I_i)$, $i = 2, ..., k$.

Finally, under (2), the likelihood is

$$\begin{aligned} &L(\mathbf{q}_1, \mathbf{q}_2, ..., \mathbf{q}_{r-1}; \mathbf{Y}) \\ &\quad = \textstyle\prod_{i=1}^k \{q_{1i}^{\sum_{j=1}^{r-1} y_{ij}} (1 - q_{1i})^{y_{ir}} q_{2i}^{\sum_{j=1}^{r-2} y_{ij}} (1 - q_{2i})^{y_{i,r-1}} \cdots q_{r-1,i}^{y_{i1}} (1 - q_{r-1,i})^{y_{i2}}\}. \end{aligned} \tag{4}$$

For the rest of this section we will assume $r = 3$, as in our example and in our simulation study, and also to simplify notation. Hence, combining the likelihood in

(4) with the prior in (3) we arrive at the Bayesian model having posterior,

$$\begin{aligned} f(\mathbf{q}_1, \mathbf{q}_2 \mid \mathbf{Y}) \;& \propto \prod_{i=1}^{k} \left\{ q_{1i}^{y_{i1}+y_{i2}} (1-q_{1i})^{y_{i3}} q_{2i}^{y_{i1}} (1-q_{2i})^{y_{i2}} \right\} \\ & \cdot q_{11}^{\gamma_1 - 1} (q_{12} - q_{11})^{\gamma_2 - 1} ... (q_{1k} - q_{1,k-1})^{\gamma_k - 1} (1 - q_{1k})^{\gamma_{k+1} - 1} \\ & \cdot q_{21}^{\delta_1 - 1} (q_{22} - q_{21})^{\delta_2 - 1} ... (q_{2k} - q_{2,k-1})^{\delta_k - 1} (1 - q_{2k})^{\delta_{k+1} - 1}, \end{aligned} \tag{5}$$

where $\gamma_i = \alpha_1 G_{01}(I_i)$ and $\delta_i = \alpha_2 G_{02}(I_i)$.

We note that Gelfand and Kuo (1991) consider the $r = 3$ case using two distinct families of priors on the η_{ij}. One arises as a product of ordered Dirichlets, the second provides a form conjugate with the likelihood in the η_{ij}. In one case F_1 and F_2 are required to be "a product Dirichlet process prior with stochastic order" though this measure is not formally defined but rather, the constraint $F_1(c) \leq F_2(c)$ is imposed at the observed dosage levels inducing a prior such that $\eta_{i1} \leq \eta_{i2}$ along with $\eta_{i1} \leq \eta_{i+1,1}$ and $\eta_{i,2} \leq \eta_{i+1,2}$ for all i. In the second case, the F_1 and F_2 are not modeled but, rather, independent priors are placed on (p_{i1}, p_{i2}, p_{i3}) directly for each i. We note that the introduction of G_1 and G_2 as in (2) provides formal Dirichlet process modeling which includes either of these cases. However, modeling F_1 and F_2 through G_1 and G_2 produces only a subset of all pairs (F_1, F_2) which are stochastically ordered. Indeed, informal checking of the product assumption can be made with the data. Criticism would be registered if at say consecutive dose levels the empirical ratio $\widehat{F}_1/\widehat{F}_2$ is not increasing.

Model fitting for (3) and (4) is carried out using Markov chain Monte Carlo. One possible implementation introduces latent variables similar to those in Gelfand and Kuo (1991). Our algorithm introduces auxiliary variables in the spirit of Besag and Green (1993) and Damien, Wakefield and Walker (1999). In the context of (5), consider positively valued random variables U_{1i}, V_{1i}, U_{2i}, V_{2i}, $i = 1,...,k$ such that the joint density of $\mathbf{q}_1$, $\mathbf{q}_2$ and $\mathbf{U}_j = (U_{j1},...,U_{jk})$, $\mathbf{V}_j = (V_{j1},...,V_{jk})$, $j = 1,2$ is given by

$$\begin{aligned} f(\mathbf{q}_1, \mathbf{q}_2, \mathbf{u}_1, \mathbf{v}_1, \mathbf{u}_2, \mathbf{v}_2 \mid \mathbf{Y}) \;\; &\propto \textstyle\prod_{i=1}^{k} (1_{(0, q_{1i}^{y_{i1}+y_{i2}})}(u_{1i}) 1_{(0,(1-q_{1i})^{y_{i3}})}(v_{1i}) \\ &\qquad\qquad 1_{(0,q_{2i}^{y_{i1}})}(u_{2i}) 1_{(0,(1-q_{2i})^{y_{i2}})}(v_{2i})) \\ &\cdot q_{11}^{\gamma_1 - 1}(q_{12} - q_{11})^{\gamma_2 - 1} ... (q_{1k} - q_{1,k-1})^{\gamma_k - 1}(1 - q_{1k})^{\gamma_{k+1}-1} \\ &\cdot q_{21}^{\delta_1 - 1}(q_{22} - q_{21})^{\delta_2 - 1} ... (q_{2k} - q_{2,k-1})^{\delta_k - 1}(1 - q_{2k})^{\delta_{k+1}-1}. \end{aligned}$$

It is clear that the marginal posterior for $\mathbf{q}_1$, $\mathbf{q}_2 \mid \mathbf{Y}$ is (5). The joint posterior only involves the prior for $\mathbf{q}_1$ and $\mathbf{q}_2$ subject to the support restrictions imposed by the uniform distributions for the U's and V's. The resulting Gibbs sampler is now routine. Updating the $\mathbf{U}$'s and $\mathbf{V}$'s given the $\mathbf{q}$'s involves uniform draws. Updating the $\mathbf{q}$'s given the $\mathbf{U}$'s and $\mathbf{V}$'s requires only draws of truncated Beta random variables. In fact, the truncating intervals can be obtained explicitly rendering the algorithm very efficient.

Prediction at a new dose d_0 is straightforward and is done after fitting the model. Introducing d_0 appropriately within the ordered dose levels introduces a corresponding $q_{\ell 0}$ into each $\mathbf{q}_\ell$. In fact, from (3) the conditional distribution of $q_{\ell 0} \mid \mathbf{q}_\ell$ is immediately a rescaled Beta distribution. Hence, a Monte Carlo integration for the predictive distribution for $q_{\ell 0}$ arises as the mixture distribution which is the average of these conditional distributions over the posterior sample of $\mathbf{q}_\ell$'s.

Finally, the inversion problem is more difficult here than in the previous section.

Now, an unknown d_0 has associated unobserved $q_{\ell 0}$'s which are not functions of d_0 but rather have specified (Beta) distributions given d_0. Now, we require a prior on $(q_{10}, q_{20}, ..., q_{r-1,0}, d_0)$. In fact, given $\mathbf{d} = (d_1, ..., d_k)$ we must specify a prior on $(\mathbf{q}_1, q_{10}, \mathbf{q}_2, q_{20}, ..., \mathbf{q}_{r-1}, q_{r-1,0}, d_0)$. At first, it seems natural to proceed as in the prediction case, i.e., given d_0, extend each $\mathbf{q}_\ell$ with $q_{\ell 0}$ to a higher dimensional Dirichlet distribution analogous to (3) and then add a prior on d_0. Updating of the q's would then be done as above. However, updating d_0 reveals the difficulty. The current d_0 positions the $q_{\ell 0}$'s relative to the other q's. But then, given the $q_{\ell 0}$'s and q's, the new d_0 must be positioned between the same observed doses as the current ones. Hence, for every iteration each $q_{\ell 0}$ will remain between the same pair of q's; d_0 will remain between the same pair of d's. Only a portion of the posteriors for the $q_{\ell 0}$'s and for d_0 will be seen in the Gibbs sampling.

Instead, we propose to update the block $(q_{10}, q_{20}, ..., q_{r-1,0}, d_0)$ all at once, updating the $\mathbf{q}_\ell$ as before. The full conditional distribution for $(q_{10}, q_{20}, ..., q_{r-1,0}, d_0)$ has an essentially intractable form. The contribution from the likelihood depends upon which interval d_0 falls in; the contribution from the prior introduces d_0 through powers involving $G_{0\ell}(\log d_0)$. We have found it easiest to discretize this full conditional in order to directly sample it, in the spirit of Ritter and Tanner (1992). Though considerable function evaluation is needed, in this way we can avoid Metropolis steps. Implicitly then, the prior on d_0 is constrained to bounded support. In fact, we use a uniform prior over this support, yielding a discrete uniform over the grid of d_0 values.

(Table 3 here)

Turning to the analysis of the data in Table 1, we take $G_{0\ell} = N(5.4, 1)$, $\ell = 1,2$. These distributions have mean roughly at the center of the data with large variance using a calculation similar to that in section 2. This illustrative data analysis and the simulation study of the next section both use some rough information from the data to specify the prior. However, subjective prior specification is implementable with, for instance, information only on the center and the range of dose values. We also set the precision parameters $\alpha_1 = \alpha_2 = 1$. This value is widely used in the literature. Moreover, we have conducted a sensitivity analysis for α_ℓ using $\alpha_\ell = 0.1$, 1, and 10, $\ell = 1,2$. In general, we obtain similar results under $\alpha_\ell = 1$ and 10 while the posteriors under $\alpha_\ell = 0.1$ depict more variability. Such results are in accordance with the role of α_ℓ in the Dirichlet process prior $DP(\alpha_\ell G_{0\ell})$ which is to control the amount of *variability* around the base distribution $G_{0\ell}$. As an illustration, in Figure 1 we plot point and 95% pointwise interval posterior estimates for $\eta_{02} = F_2(\log d_0)$ as a function of new dose d_0.

(Figure 1 here)

Under the above prior specification, posterior summaries for the p_{ij} are included in Table 3. The posterior means for η_{01} and η_{02} are included in Figure 2.

(Figure 2 here)

Finally, we considered the inversion problem at the same two choices of $\mathbf{Y}_0$ as in section 2. Using a discrete uniform prior over the interval $(10, 5000)$ we obtain point (posterior median) and 95% equal tail interval estimates as shown in Table 4.

(Table 4 here)

4 Simulation study

In this section, through simulated data, we study the performance of the nonparametric model when the parametric model is true and demonstrate its superiority when the actual model that generates the data is nonstandard. The two simulated examples are described in detail in sections 4.1 and 4.2. To make the study directly pertinent to the analyses of sections 2 and 3, we choose a setting similar to the one for the data in Table 1, taking $r = 3$, $k = 7$ throughout and the same values for dose level. We also use the same priors given in sections 2 and 3 for the parametric and nonparametric model, respectively. Finally, we consider the effect of the sample sizes n_i generating two sets of data for each case, one with sample sizes $n_i^{(1)}$ as in Table 1 and one with smaller sample sizes an order of magnitude smaller, i.e., $n_i^{(2)} = [n_i^{(1)}/10]$.

4.1 Simulated Data from the Parametric Model

For the first simulation, we employ the probabilities postulated by the parametric model in (1). We compute them using $\beta_{11} = -8$, $\beta_{12} = -5$, $\beta_{21} = 1$ and $\beta_{22} = 0.7$ which are roughly the posterior point estimates from Table 2. Then the two sets of data are generated from Multinomial distributions with these probabilities and sample sizes $n_i^{(1)}$ and $n_i^{(2)}$. Figure 3 contains point and 95% pointwise interval posterior estimates of $F_1(\log d_0)$ and $F_2(\log d_0)$, for a grid of values of new dose d_0, for both data sets and under both models. In all cases, wider interval estimates emerge under the nonparametric model. In fact, this is the main difference for the smaller sample sizes, the point estimates being similar with the nonparametric point estimates following the data more closely.

(Figure 3 here)

4.2 Simulated Data from a Nonstandard Model

Here we consider a setting under which model (1) is anticipated to perform poorly. In the spirit of the formulation in (2), we set $p_{i1} = F_1(\log d_i) = G_1(\log d_i)G_2(\log d_i)$ and $p_{i1} + p_{i2} = F_2(\log d_i) = G_1(\log d_i)$ with G_1 and G_2 taken to both be two component mixtures of normal distributions. More precisely, $G_1 = 0.5N(3.5, (0.25)^2)$ $+ 0.5N(5.85, (0.2)^2)$ and $G_2 = 0.3N(3.6, (0.3)^2) + 0.7N(5.99, (0.5)^2)$. Based on the probabilities p_{i1} and p_{i2} as defined above, we generate two data sets from Multinomial distributions with sample sizes $n_i^{(1)}$ and $n_i^{(2)}$. Figure 4 shows the simulated data and provides posterior estimates exactly analogous to these of Figure 3. The superiority of the nonparametric model is evident recovering the functional forms of F_1 and F_2 very succesfully with the predictive accuracy increasing with increasing sample sizes. The parametric model again leads to tighter interval estimates but here, as anticipated, provides inaccurate inference for F_1 and F_2.

(Figure 4 here)

5 Comparison of the Analyses and Related Remarks

We draw further comparisons between the two fitted models supplementing the discussion of section 4 with the results of sections 2 and 3. These comparisons are only sensible with regard to estimation of probabilities and dose inversion. In Table 3 we see generally good agreement in the estimation of the p_{ij}'s. As would be expected, interval estimates grow wider as dose increases. Also, as expected, interval estimates are tighter for the parametric case. The large number of blood cells measured at each dose enables tight estimates of the β_{1j} and β_{2j}, hence for the p_{ij}.

Turning to Figure 2, we again see good agreement between the models. Notice, however, that the more flexible nonparametric model more closely follows the observed $\widehat{\eta}_{ij}$, $j = 1,2$. This may suggest overfitting with a resultant trade-off in larger predictive variability. For instance, for the inversion problem, looking at the validation case (column 2 of Table 4), while both intervals are properly centered (again true dose is 100), the nonparametric one is substantially wider.

The second inversion example (column 1 of Table 4) suggests, from Figure 2, a dose beyond 500 since $\widehat{\eta}_{01} = 0.1302$ and $\widehat{\eta}_{02} = 0.4602$. For this extrapolation the two models differ considerably. In fact, Figure 2 at $d = 500$ shows that the nonparametric curves tend to lie below the parametric ones. This suggests that, for high dose extrapolation, the predicted d_0 under the nonparametric model will exceed that of the parametric one. Indeed, this is the result in Table 4. Also note that, as expected, the prediction intervals are wider for the nonparametric model. Cancer researchers would prefer the tighter intervals associated with the parametric model but the simple linear form in (1) may be supplying incorrect centering or optimistic precision as both cases of the simulation study of section 4 illustrate.

Nonlinear parametric models provide another class of possible models. However, they may be awkward to fit and may yield wide prediction intervals, as the result from Madruga, et al. (1996) suggests. Moreover, in the absence of mechanistic knowledge about the cell aberration process, it may be difficult to pick a suitable form. The nonparametric specification avoids having to make this choice.

Finally we note an illustrative semiparametric analysis. Since $\eta_{ij} \in (0,1)$ and increases in j, it is natural to model $\{\eta_{ij}\}$ using a c.d.f. So we could set

$$\eta_{ij} = F(\log d_i + \triangle_j). \tag{6}$$

In (6), F is an unknown c.d.f. with $\triangle_j$ providing an adjustment for the j^{th} partial sum. Expression (6) implies that the η_{ij} increase in d_i. But also, since $\eta_{ij} < \eta_{i,j+1}$, we require $\triangle_j < \triangle_{j+1}$. To *identify* F and the $\triangle_j$ it is convenient to set $\triangle_1 = 0$. The resulting model is semiparametric in that the model unknowns are the arbitrary c.d.f. F, and the set of $r-1$ $\triangle_j$'s. F might be modeled through a Dirichlet process as in the previous section though a more convenient choice within our setting would be to use the mixture-of-Betas approach described in Mallick and Gelfand (1994). A limitation of (6) is that the change $\eta_{i,j+1} - \eta_{ij}$ is only captured by the shift $\triangle_{j+1} - \triangle_j$. The model in (2) provides a distinct c.d.f. for each j.

ACKNOWLEDGEMENTS

The work of the second author was done while visiting in the Department of Statistics at the University of Connecticut and was supported under a grant from FAPES 98/6062-6. The work of the first and third authors was supported in part by NSF DMS 99-71206. The authors acknowledge Carlos Pereira, Department of Statistics, University of São Paulo for introducing the problem to them and for helpful discussion.

REFERENCES

Aitchison, J. and Shen, S.M. (1980). Logistic-normal distributions: Some properties and uses. *Biometrika* **67**, 261-272.

Bender, M.A., Awa, A.A., Brooks, A.L., Evans, H.J., Groer, P.G., Littlefield, L.G., Pereira, C.A. de B., Preston, F.J. and Wachholz, B.W. (1988). Current status of cytogenetic procedures to detect and quantify previous exposures to radiation. *Mutation Research* **196**, 103-159.

Besag, J. and Green, P.J. (1993). Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society, Ser. B* **55**, 25-37.

Damien, P., Wakefield, J. and Walker, S. (1999). Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society, Ser. B* **61**, 331-344.

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

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

Frome, E.L. and DuFrain, R.J. (1986). Maximum likelihood estimation for cytogenetic dose-response curves. *Biometrics* **42**, 73-84.

Gelfand, A.E. and Kuo, L. (1991). Nonparametric Bayesian bioassay including ordered polytomous response. *Biometrika* **78**, 657-666.

Gelfand, A.E. and Sahu, S.K. (1999). Gibbs sampling, identifiability and improper priors in generalized linear mixed models. *Journal of the American Statistical Association* **94**, 247-253.

Hobert, J.P. and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association* **91**, 1461-1473.

Johnson, V.E. and Albert, J.H. (1999). *Ordinal Data Modeling.* New York: Springer.

Madruga, M.R., Pereira, C.A. de B. and Rabello-Gay, M.N. (1994). Bayesian dosimetry: Radiation dose versus frequencies of cells with aberrations. *Environmetrics* **5**, 47-56.

Madruga, M.R., Ochi-Lohmann, T.H., Okazaki, K., Pereira, C.A. de B. and Rabello-Gay, M.N. (1996). Bayesian dosimetry II: Credibility intervals for radiation dose. *Environmetrics* **7**, 325-331.

Mallick, B.K. and Gelfand, A.E. (1994). Generalized linear models with unknown link function. *Biometrika* **81**, 237-245.

Mukhopadhyay, S. (2000). Bayesian nonparametric inference on the dose level with specified response rate. *Biometrics* **56**, 220-226.

Osborne, D. (1991). Statistical calibration: A review. *International Statistical Review* **59**, 309-336.

Pereira, C.A. de B. and Pericchi, L.R. (1990). Analysis of diagnosability. *Applied Statistics* **39**, 189-204.

Ritter, C. and Tanner, M.A. (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the griddy - Gibbs sampler. *Journal of the American Statistical Association* **87**, 861-868.

Spiegelhalter, D.J., Thomas, A., Best, N. and Gilks, W.R. (1995). *BUGS: Bayesian Inference Using Gibbs Sampling, Version 0.50.* Medical Research Council, Biostatistics Unit, Cambridge, U.K.

Table 1: Observed frequencies for binucleated cells from healthy older subjects. y_1 denotes at least two MN, y_2 exactly one MN, y_3 0 MN

i	Dose (cGy)	y_{i1}	y_{i2}	y_{i3}	$\widehat{\eta}_{i1}$	$\widehat{\eta}_{i2}$
1	20	8	41	989	0.0077	0.0472
2	50	14	56	933	0.0140	0.0698
3	100	32	114	939	0.0295	0.1346
4	200	67	176	794	0.0646	0.2343
5	300	59	209	683	0.0620	0.2818
6	400	107	256	742	0.0968	0.3285
7	500	143	327	771	0.1152	0.3787

Table 2: For the data in Table 1, posterior summary for parametric model of section 2

	mean	**sd**	**2.5%**	**median**	**97.5%**
β_{11}	-8.081	0.3936	-8.883	-8.071	-7.339
β_{12}	-5.662	0.2318	-6.143	-5.661	-5.214
β_{21}	1.025	0.0688	0.893	1.023	1.165
β_{22}	0.775	0.0418	0.694	0.774	0.861

Table 3: For the data in Table 1, posterior mean and 95% equal tail interval estimates for the p_{ij} under the models in sections 2 and 3

p_{i1}

Dose	Parametric	Nonparametric
20	0.0065	0.0108
	(0.0043, 0.0091)	(0.0055, 0.0156)
50	0.0158	0.0141
	(0.0120, 0.0200)	(0.0093, 0.0199)
100	0.0300	0.0309
	(0.0251, 0.0353)	(0.0230, 0.0391)
200	0.0551	0.0607
	(0.0494. 0.0611)	(0.0496, 0.0737)
300	0.0766	0.0737
	(0.0699, 0.0837)	(0.0598, 0.0893)
400	0.0955	0.0985
	(0.0866, 0.1049)	(0.0822, 0.1147)
500	0.1122	0.1129
	(0.1005, 0.1246)	(0.0982, 0.1297)

p_{i2}

Dose	Parametric	Nonparametric
20	0.0341	0.0429
	(0.0273, 0.0416)	(0.0299, 0.0561)
50	0.0662	0.0522
	(0.0574, 0.0754)	(0.0409, 0.0665)
100	0.1063	0.1037
	(0.0971, 0.1158)	(0.0873, 0.1222)
200	0.1643	0.1780
	(0.1553, 0.1736)	(0.1565, 0.2017)
300	0.2066	0.2102
	(0.1958, 0.2175)	(0.1836, 0.2392)
400	0.2395	0.2341
	(0.2258, 0.2535)	(0.2082, 0.2615)
500	0.2661	0.2594
	(0.2493, 0.2835)	(0.2367, 0.2832)

p_{i3}

Dose	Parametric	Nonparametric
20	0.9594	0.9464
	(0.9512, 0.9664)	(0.9311, 0.9622)
50	0.9181	0.9337
	(0.9079, 0.9272)	(0.9160, 0.9475)
100	0.8636	0.8654
	(0.8531, 0.8735)	(0.8437, 0.8844)
200	0.7806	0.7613
	(0.7701, 0.7907)	(0.7330, 0.7892)
300	0.7168	0.7161
	(0.7044, 0.7290)	(0.6805, 0.7496)
400	0.6650	0.6674
	(0.6496, 0.6801)	(0.6294, 0.7022)
500	0.6217	0.6277
	(0.6033, 0.6397)	(0.5978, 0.6558)

Table 4: Posterior summaries for the dose inversion problem using the models of sections 2 and 3 for the data in Table 1

	Observed $\mathbf{Y}_0$	$\mathbf{Y}_0 = (316, 801, 1310)$	$\mathbf{Y}_0 = (32, 114, 939)$
	Prior on d_0	Posterior Point (Interval Estimate)	
Parametric Model	Lognormal(4.96,9)	751.2 (647.7, 873.1)	96.3 (74.5, 121.0)
Nonparametric Model	Discrete Uniform	865.1 (557.9, 1670.2)	74.3 (11.0, 169.6)

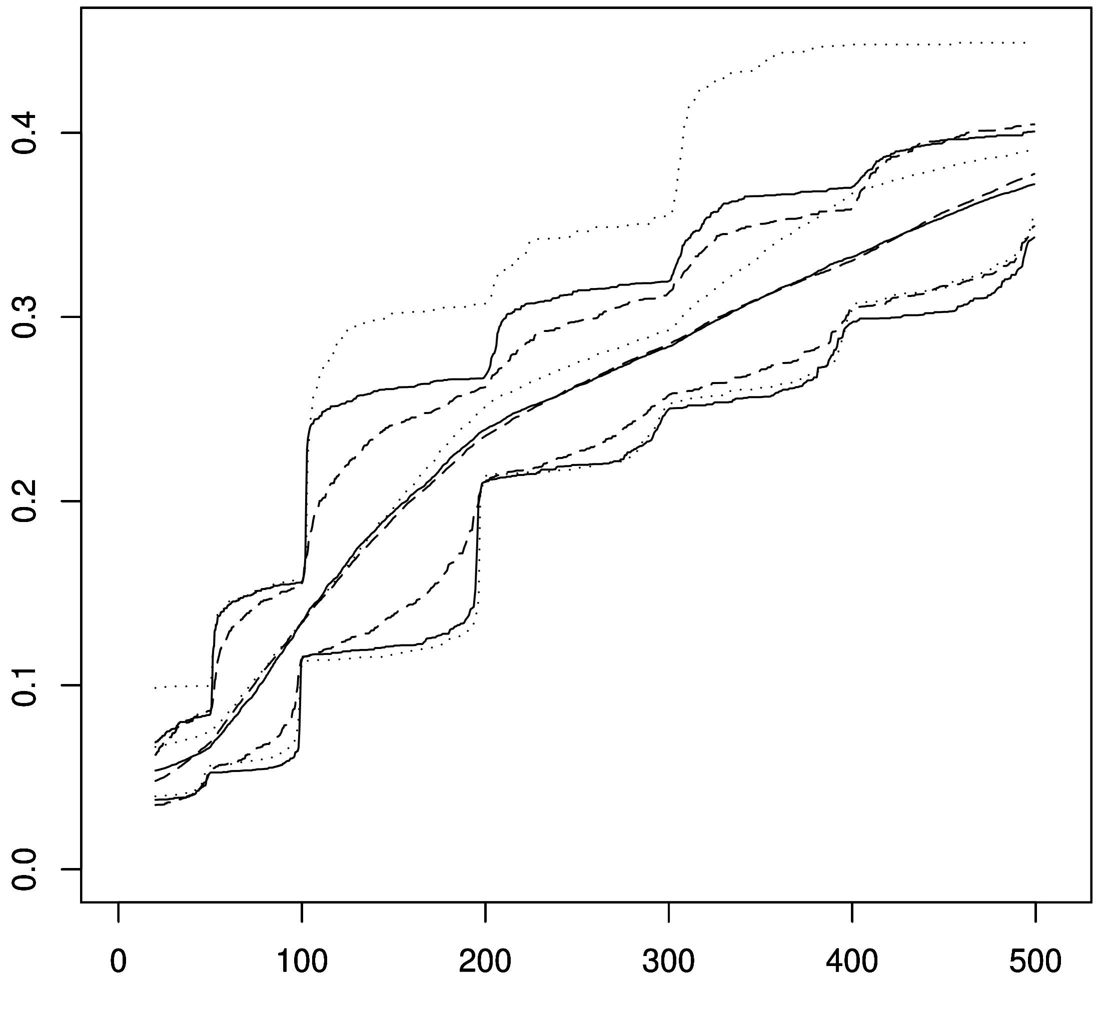

Figure 1: For the data in Table 1, point and 95% pointwise interval posterior estimates for the probability of at least one MN vs d_0 under $\alpha_1 = \alpha_2 = 0.1$ (dotted lines), 1 (solid lines) and 10 (dashed lines).

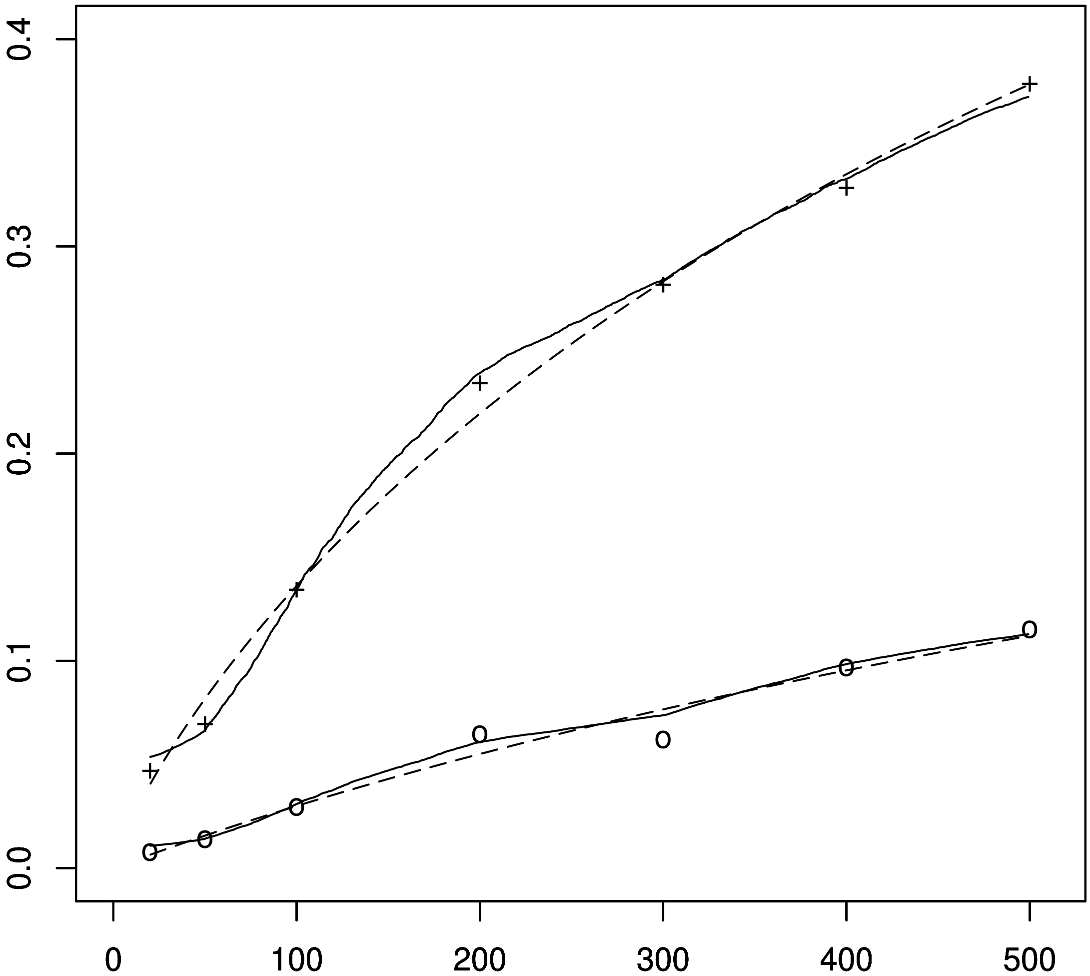

Figure 2: For the data in Table 1, prediction of the probability of two or more MN (lower curves) and at least one MN (upper curves) vs d_0. The solid lines correspond to the nonparametric model and the dashed lines to the parametric model. The observed $\widehat{\eta}_{i1}$ are denoted by "o" and the observed $\widehat{\eta}_{i2}$ by "+".

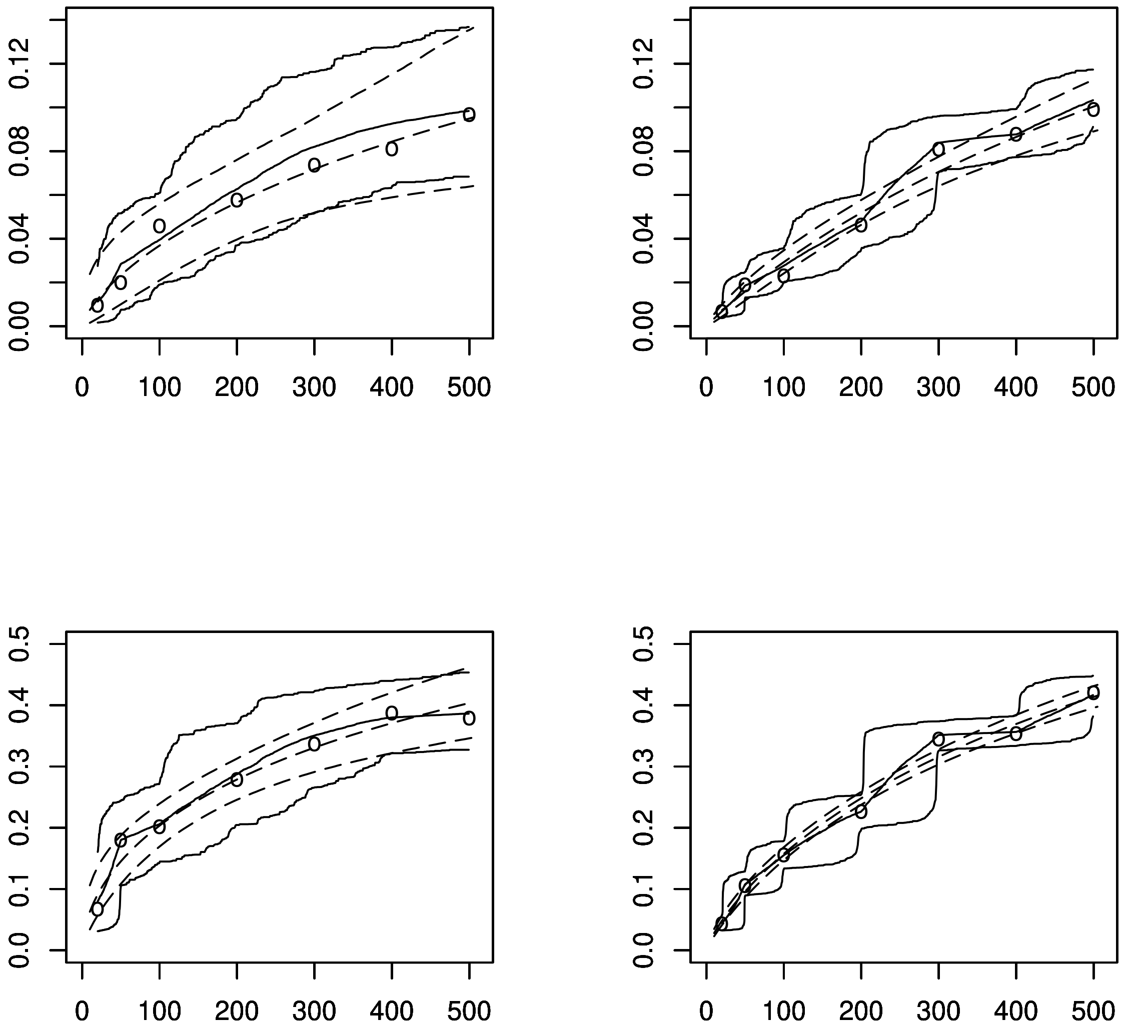

Figure 3: For the simulated data of section 4.1, posterior inference for F_1 (upper panels) and F_2 (lower panels) under the parametric (dashed lines) and nonparametric (solid lines) model. "o" denotes the observed data. The left and right panels correspond to the data set with the smaller and large sample sizes, respectively.

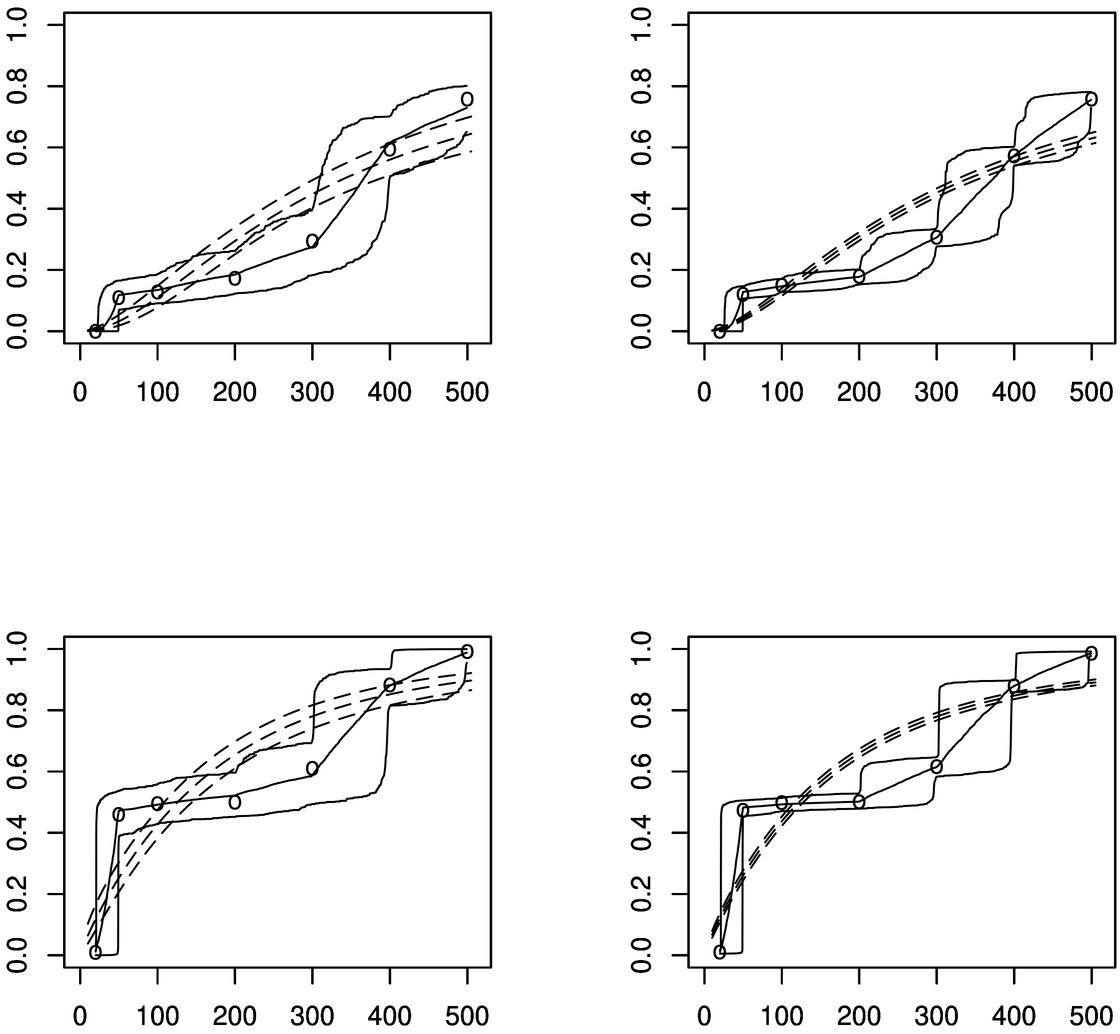

Figure 4: For the simulated data of section 4.2, posterior inference for F_1 (upper panels) and F_2 (lower panels) under the parametric (dashed lines) and nonparametric (solid lines) model. "o" denotes the observed data. The left and right panels correspond to the data set with the smaller and large sample sizes, respectively.