RT-MAE 2001-01

A NONPARAMETRIC BAYESIAN MODELING APPROACH FOR CYTOGENETIC DOSIMETRY

by

Athanasios Kottas, Márcia D. Branco
and
Alan E. Gelfand

Palavras-Chave: Auxiliary variables; Dirichlet process; dose-response; logistic regression; polytomous response.
Classificação AMS: 62G07, 62F15, 62J12.
(AMS Classification)

- Janeiro de 2001 -

A Nonparametric Bayesian Modeling Approach for Cytogenetic Dosimetry

Athanasios Kottas, Márcia D. Branco and Alan E. Gelfand*

Abstract

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 so-called calibration 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 natural 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, offering comparison with a parametric model. We examine a dataset modeling blood cultures exposed to radiation where classification is with regard to number of micronuclei per cell.

*A. Kottas is Visiting Assistant Professor in the Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708-0251, USA, Márcia D. Branco is Assistant Professor in the Departamento de Estatística, IME, Universidade de São Paulo, Cxa Postal 66281, 05315-970, São Paulo, SP, Brasil and A.E. Gelfand is Professor in the Department of Statistics, University of Connecticut, Storrs, CT 06269-3120, USA. 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.

KEY WORDS: Auxiliary variables; Dirichlet process; dose-response; logistic regression; polytomous response.

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. That 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. As noted, human exposure is difficult to assess accurately. Moreover, even with reasonably reliable exposure measurements, typically, uncomfortable extrapolation arises. Measurements are at lower doses while interest is in the relationship at higher levels. A very thorough, readable discussion is given in Bender, et al. (1988).

In the in vitro setting, which we confine ourselves to, experimentation is more straightforward. 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.

In fact, it is often a count which, in the literature, 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. See, e.g., Frome and DuFrain (1986) for a statistical development and related references. The inversion or statistical calibration problem is straightforward in this case. See, e.g., Osborne (1991) for a general review.

In some cases, a more natural response is a categorical classification. This is our focus

here. The case of binary response, i.e., 1 indicates response observed, 0 not observed, yields the well-studied, standard bioassay problem. In the cytogenetic dosimetry literature, the polytomous response case has been examined, as in, e.g., Madruga, et al. (1994) and Madruga, et al. (1996) but exclusively under parametric models.

We propose a flexible nonparametric model to approach 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 but limited work in this regard. The Bayesian perspective is particularly helpful for the inversion problem under categorical response. That is, simultaneous inversion of a set of response curves, one for each classification, to obtain a common dose, is not a well defined problem. The Bayesian approach, 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 simpler binary response case where only one response curve is involved.

In the standard Bayesian modeling for this context, a multinomial model for the categorical response with a conjugate Dirichlet distribution on the probabilities is assumed. Following Aitchison and Shen (1980), the log ratio transformed probabilities follow a multivariate normal distribution, enabling a posterior which is multivariate normal on the transformed scale, with mean and variance as developed in Pereira and Pericchi (1990). Unfortunately, such modeling can not handle the prediction or inversion which we seek. In order to do so, Madruga, et al. (1994) introduce a parametric model for the log ratio transformed probabilities, as a function of dose, but their resultant treatment of the inversion problem is ad hoc.

We note that in 1986, a scientific committee was established, at the request of the National Cancer Institute, which undertook the assessment of the status of cytogenetic procedures to detect and quantify previous exposures to radiation. The aforementioned paper by Bender, et al. (1988) summarizes this committee's findings. Interestingly, the committee takes an 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". This predates the boom in simulation-based model fitting, which has resulted in a dramatic increase in the use of Bayesian modeling and inference. With the ability to investigate more general semiparametric and nonparametric models, the committee's stance seems even more appropriate today.

In section 2 we present a motivating data set, taken from Madruga, et al. (1996) and 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. Finally, in section 4 we offer 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). 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. The data are provided in Table 1. Also given are the sample estimates of at least two micronuclei, i.e., $\widehat{\eta}_{i1} = y_{i1}/(y_{i1} + y_{i2} + y_{i3})$ and at least one micronuclei, i.e., $\widehat{\eta}_{i2} = (y_{i1} + y_{i2})/(y_{i1} + y_{i2} + y_{i3})$.

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. For the data in Table 1, $k = 7$ and $r = 3$.

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, be able to predict $\mathbf{Y}_0$. Inversely, we would like to estimate an unobserved dose level given an observed $\mathbf{Y}_0$. The latter two problems presume some sort of 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

$$\log \frac{p_{ij}}{p_{ir}} = \alpha_j + \beta_j \log d_i. \tag{1}$$

More complicated, nonlinear forms are discussed in Madruga, et al. (1996). In fact, they work with the larger dataset and introduce an effect for whether the cells are mononucleated or binucleated.

Bayesian inference under (1) is now standard. With a flat prior on the (α_j, β_j), 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, e.g., Gelfand and Sahu (1999) and further references therein. Model fitting is routine using the BUGS software (Spiegelhalter, et al. 1995).

More generally, a bivariate normal prior could be introduced for (α_j, β_j). If sensible, the (α_j, β_j) might be assumed i.i.d., adding hyperparameters and a hyperprior, creating a hierarchical model. Conditions for posterior propriety are known, following, e.g., Hobert and Casella (1996).

For the data in Table 1 we adopt the flat prior assumption, presenting posterior summaries for the α_j and β_j 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_1 and β_2 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 prominent 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 Figures 1 and 2, respectively. 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.

Madruga, et al. (1996) consider the inversion problem for a healthy older subject showing $\mathbf{Y}_0 = (316, 801, 1310)$. Relative to Table 1, a d_0 larger than 500 is clearly suggested, yielding an extrapolation problem. Using their model with an ad hoc inversion, (815, 1210) is obtained as a roughly 95% credible interval for the unobserved dose. We also attempt to validate our inversion taking $\mathbf{Y}_0 = (32, 114, 939)$, in fact, the observed data at $d = 100$. Regardless, 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. Using an illustrative normal prior on $\log d_0$ we obtain the point (posterior median) and 95% equal tail interval estimate for d_0 provided in Table 4. This prior is centered at the average log dose for our sample, 4.96, with variance 9 which is very large in our situation. 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. Due to the large number of blood cells observed at d_0, the data essentially overwhelms this prior. Hence, we have prior robustness though our model may not be very good. 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 α_j and β_j, 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 α, β and $\log d_0$ is proportional to $\left\{\prod_{i=1}^{k} e^{(\alpha+\beta \log d_i)y_i}/(1+e^{\alpha+\beta \log d_i})^{n_i}\right\} e^{(\alpha+\beta \log d_0)y_0}/(1+e^{\alpha+\beta \log d_0})^{n_0}$ Reparametrizing to α, β and $z_0 = \beta \log d_0$ and then integrating over z_0 yields the form $c(\alpha)\beta^{-1}\prod_{i=1}^{k} e^{(\alpha+\beta \log d_i)y_i}/(1+e^{\alpha+\beta \log d_i})^{n_i}$ which is not integrable with respect to β.

3 A Fully Nonparametric Approach

Now, we propose a fully nonparametric model which is also straightforward to fit. Again, 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 $\eta_{ij} < \eta_{i,j+1}$.

The G_ℓ's are arbitrary and unknown and hence random. They are introduced because we propose to model them as independent random functions. It is easier to work with an independent set of unknown random distributions than to model the F_j directly. To simplify notation, we let $G_\ell(\log d_i) = q_{\ell i}$.

We propose that each G_ℓ be modeled as a realization from a Dirichlet process (Ferguson, 1973), 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. That is, for any partition of R^1, say $(B_1, B_2, ..., B_t)$, the random vector

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

Furthermore, the G_ℓ are taken to be independent.

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.,

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

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 readily assembled, yielding the convenient form

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

For the balance of this section we will assume $r = 3$, as in our example, and to keep notation a bit simpler. Hence, combining the likelihood in (4) with the prior in (3) we arrive at the Bayesian model having posterior,

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

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.

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). In the context of our present illustration, let $W_{1,i,i'}$ denote the unobserved number of cells at dose level i which would have produced at least one MN at dose level i' but not at dose level $i'-1$. Similarly, let $W_{2,i,i'}$ denote the unobserved number of cells at dose level i which would have produced at least two MN at dose level i' but not at dose level $i'-1$. Define $W_{1,i,k+1}$ and $W_{2,i,k+1}$ in the obvious way. Then $\mathbf{W}_{1i} \mid \mathbf{q}_1 \sim Mult(n_i, q_{11}, (q_{12} - q_{11}), ..., (1 - q_{1k}))$ and $\mathbf{W}_{2i} \mid \mathbf{q}_2 \sim Mult(y_{i1} + y_{i2}; q_{21}, (q_{22} - q_{21}), ..., (1 - q_{2k}))$, subject to $\sum_{i'=1}^{i} W_{1,i,i'} = y_{i1} + y_{i2}$ and $\sum_{i'=1}^{i} W_{2,i,i'} = y_{i1}$. Inspection of (5) reveals that the "full data" likelihood, i.e., given the W's is conjugate with the prior. Hence, a routine Gibbs sampler can be run to update the q's given the W's and the W's given the q's.

An alternative implementation, which requires less bookkeeping, is to introduce auxiliary variables in the spirit of Besag and Green (1993) and Damien, et al. (1999). Again, in the context of (5), let U_{1i} be independent, $U_{1i} \sim U(0, q_{1i}^{y_{i1}+y_{i2}})$ and let V_{1i} be independent, $V_{1i} \sim U(0, (1 - q_{1i})^{y_{i3}})$. Similarly, let $U_{2i} \sim U(0, q_{2i}^{y_{i1}})$ and $V_{2i} \sim U(0, (1 - q_{2i})^{y_{i2}})$. These choices provide a Gibbs sampler which, again, is routine. Updating the U's and V's given the q's involves uniform draws. Updating the q's given the U's and V 's requires only draws of truncated Beta random variables. It is evident that the auxiliary variables implementation introduces far fewer additional variables to the model than the latent variables implementation. We employed the auxiliary variables sampler in the analysis of the data in Table 1.

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 $\mathbf{q}_\ell$. The predictive distribution for $q_{\ell 0}$ arises as the mixture distribution which is the average of the full conditional distributions for $q_{\ell 0}$ at each posterior sample.

Finally, the inversion problem is a bit 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 one. The Gibbs sampler becomes trapped in a subset of the entire parameter space.

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 $(q_{10}, q_{20}, ..., q_{r-1,0}, d_0)$ is very awkward to work with. The contribution from the likelihood depends upon where d_0 falls; 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.

Turning to the analysis of the data in Table 1, we took $\alpha_1 = \alpha_2 = 1$ with $G_{01} = N(4.5, 1)$ and $G_{02} = N(4.5, 1)$. These distributions have mean roughly at the center of the data with large variance using calculation similar to that in section 2. Posterior summaries for the p_{ij} are included in Table 3. The posterior means for η_{01} and η_{02} are included in Figures 1 and 2, respectively. Again interval estimates are routine from the sampling but are not shown.

Finally, we considered the inversion problem at the same two choices of $\mathbf{Y}_0$ as above. 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.

4 Comparison of the Analyses and Related Remarks

We draw some comparisons between the two fitted models. 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 enable tight estimates of the α_j and β_j, hence for the p_{ij}.

Turning to Figures 1 and 2, we again see good agreement between the models. Notice, however, that the more flexible nonparametric model more closely follows the observed $\widehat{\eta}_{ij}$. This may imply overfitting with a resultant trade-off in larger predictive variability. For the inversion problem, looking at the validation illustration (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 illustration (column 1 of Table 4) suggests a dose beyond 500 since $\widehat{\eta}_{01} = .1302$ and $\widehat{\eta}_{02} = .4602$. For this extrapolation the two models differ considerably. In fact, Figures 1 and 2 at $d = 500$ show that the nonparametric curve lies below the parametric one. This suggests that 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 providing incorrect centering.

Nonlinear parametric models provide another class of possible models. However, they may be awkward to fit and may yield wide prediction intervals, as the ad hoc 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 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

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

In (6), F is an unknown c.d.f. with Δ_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 $\Delta_j < \Delta_{j+1}$. To *identify* F and the Δ_j it is convenient to set $\Delta_1 = 0$. The resulting model is semiparametric in that the model unknowns are F, an arbitrary c.d.f., and the set of $r-1$ Δ_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 $\Delta_{j+1} - \Delta_j$. The model in (2) provides a distinct c.d.f. for each j.

References

[1] Aitchison, J. and Shen, S.M. (1980). Logistic-normal Distributions: Some Properties and Uses. *Biometrika,* 67, 261-272.

[2] 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.

[3] Besag, J. and Green, P.J. (1993). Spatial Statistics and Bayesian Computation. *J.R. Statist. Soc. B,* 55, 25-37.

[4] Damien, P., Wakefield, J. and Walker, S. (1999). Gibbs Sampling for Bayesian Non-conjugate and Hierarchical Models by Using Auxiliary Variables. *J.R. Statist. Soc. B,* 61, 331-344.

[5] Ferguson, T.S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics,* 1, 209-230.

[6] Frome, E.L. and DuFrain, R.J. (1986). Maximum Likelihood Estimation for Cytogenetic Dose-Response Curves. *Biometrics,* 42, 73-84.

[7] Gelfand, A.E. and Kuo, L. (1991). Nonparametric Bayesian Bioassay Including Ordered Polytomous Response. *Biometrika,* 78, 657-666.

[8] Gelfand, A.E. and Sahu, S.K. (1999). Gibbs Sampling, Identifiability and Improper Priors in Generalized Linear Mixed Models. *J. Amer. Stat. Assoc.,* 94, 247-253.

[9] Hobert, J.P. and Casella, G. (1996). The Effect of Improper Priors on Gibbs Sampling in Hierarchical Linear Mixed Models. *J. Amer. Stat. Assoc.,* 91, 1461-1473.

[10] 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.

[11] 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.

[12] Mallick, B.K. and Gelfand, A.E. (1994). Generalized Linear Models with Unknown Link Function. *Biometrika*, 81, 237-245.

[13] Mukhopadhyay, S. (2000). Bayesian Nonparametric Inference on the Dose Level with Specified Response Rate. *Biometrics*, 56, 220-226.

[14] Osborne, D. (1991). Statistical Calibration: A Review. *International Statistical Review*, 59, 309-336.

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

[16] Ritter, C. and Tanner, M.A. (1992). Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy - Gibbs Sampler. *J. Amer. Stat. Assoc.*, 87, 861-868.

[17] 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	.0077	.0472
2	50	14	56	933	.0140	.0698
3	100	32	114	939	.0295	.1346
4	200	67	176	794	.0646	.2343
5	300	59	209	683	.0620	.2818
6	400	107	256	742	.0968	.3285
7	500	143	327	771	.1152	.3787

Table 2: Posterior summary for parametric model of section 2

	mean	**sd**	**2.5%**	**median**	**97.5%**
α_1	-8.081	0.3936	-8.883	-8.071	-7.339
α_2	-5.662	0.2318	-6.143	-5.661	-5.214
β_1	1.025	0.06876	0.893	1.023	1.165
β_2	0.7746	0.04177	0.694	0.774	0.8611

Table 3: 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	0.0087
	(0.0043, 0.0091)	(0.0038, 0.0133)
50	0.0158	0.0138
	(0.0120, 0.0200)	(0.0086, 0.0198)
100	0.0300	0.0312
	(0.0251, 0.0353)	(0.0223, 0.0392)
200	0.0551	0.0626
	(0.0494. 0.0611)	(0.0501, 0.0748)
300	0.0766	0.0743
	(0.0699, 0.0837)	(0.0606, 0.0887)
400	0.0955	0.0963
	(0.0866, 0.1049)	(0.0808, 0.1126)
500	0.1122	0.1055
	(0.1005, 0.1246)	(0.0917, 0.1220)

p_{i2}

Dose	Parametric	Nonparametric
20	0.0341	0.0411
	(0.0273, 0.0416)	(0.0299, 0.0535)
50	0.0662	0.0538
	(0.0574, 0.0754)	(0.0414, 0.0686)
100	0.1063	0.1034
	(0.0971, 0.1158)	(0.0864, 0.1215)
200	0.1643	0.1777
	(0.1553, 0.1736)	(0.1546, 0.2008)
300	0.2066	0.2062
	(0.1958, 0.2175)	(0.1779, 0.2369)
400	0.2395	0.2409
	(0.2258, 0.2535)	(0.2099, 0.2686)
500	0.2661	0.2601
	(0.2493, 0.2835)	(0.2359, 0.2861)

	p_{i3}	
Dose	Parametric	Nonparametric
20	0.9594	0.9503
	(0.9512, 0.9664)	(0.9357, 0.9635)
50	0.9181	0.9324
	(0.9079, 0.9272)	(0.9133, 0.9475)
100	0.8636	0.8653
	(0.8531, 0.8735)	(0.8440, 0.8861)
200	0.7806	0.7597
	(0.7701, 0.7907)	(0.7309, 0.7882)
300	0.7168	0.7195
	(0.7044, 0.7290)	(0.6809, 0.7543)
400	0.6650	0.6628
	(0.6496, 0.6801)	(0.6287, 0.7030)
500	0.6217	0.6344
	(0.6033, 0.6397)	(0.6030, 0.6631)

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

	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	733.9(533.0, 1305.1)	74.3(11.0, 169.6)

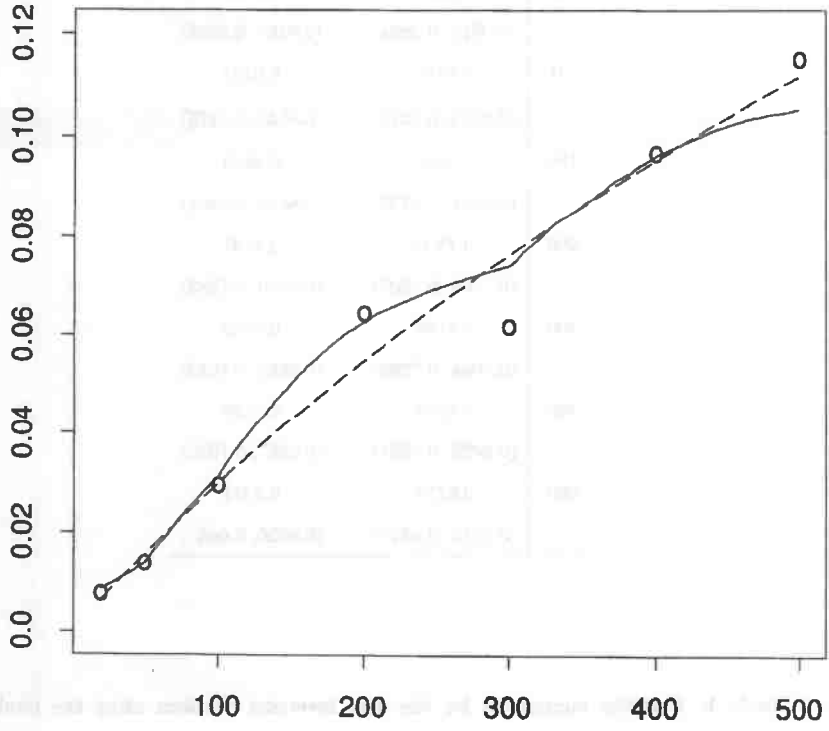

Figure 1: Prediction of the probability of two or more MN vs d_0. The solid line corresponds to the nonparametric model and the dashed line to the parametric model. The observed $\widehat{\eta}_{i1}$ are denoted by "o".

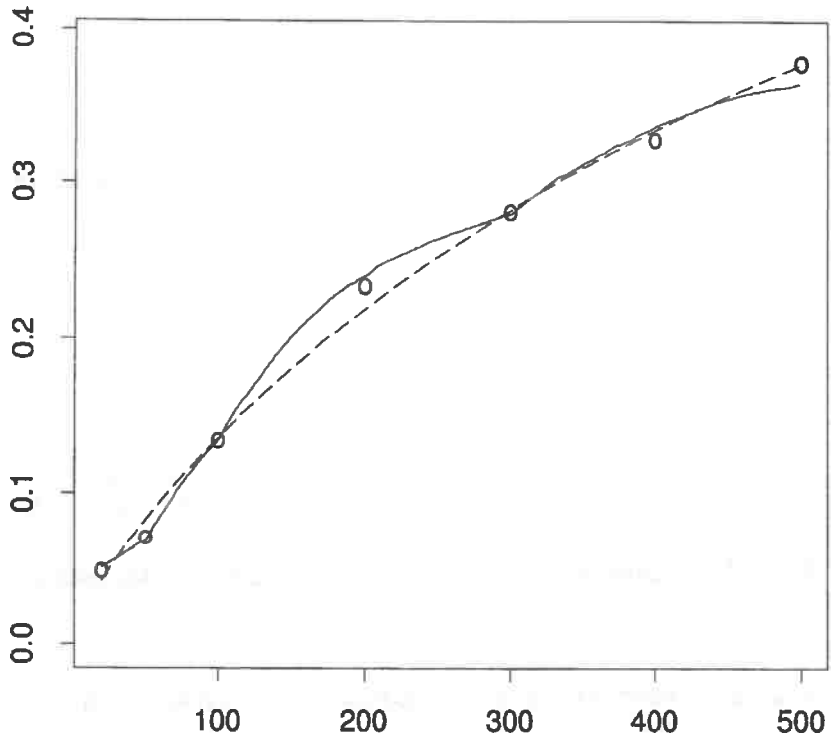

Figure 2: Prediction of the probability of at least one MN vs d_0 under the nonparametric model (solid line) and the parametric model (dashed line). The observed $\widehat{\eta}_{i2}$ are denoted by "o".

ÚLTIMOS RELATÓRIOS TÉCNICOS PUBLICADOS

2000-1 - **POPOV, S.Y., MACHADO, F.P.** One-dimensional branching random walk in a periodic random environment. 2000. 10p. (RT-MAE-2000-1)

2000-2 - **BORGES, W.S., HO L.L.; TURNES, O.** An analysis of taguchi's on-line quality monitoring procedure for attributes with diagnosis errors. 2000. 23p.(RT-MAE-2000-2)

2000-3 - **CORDEIRO, G.M., FERRARI, S.L.P., URIBE-OPAZO, M.A.** Bartlett-type corrections for two-parameter exponential family models. 2000. 21p. (RT-MAE-2000-3)

2000-4 - **SCARPA, O.** On coincidence of critical parameters in Poisson percolation model. 2000. 8p. (RT-MAE-2000-4)

2000-5 - **CANCHO, V.G., BOLFARINE, H.** Modelling the presence of immunes by using the exponentiated-weibull model. 2000. 18p. (RT-MAE-2000-5)

2000-6 - **BORGES, W., DIMITROV, B., KHALIL, Z., KOLEV, N.** System's performance under mixed minimal and imperfect repair maintenance strategies. 2000. 17p. (RT-MAE-2000-6)

2000-7 - **BARROSO, L.P., CORDEIRO, G.M., VASCONCELLOS, K.L.P.** Second-Order Asymptotics for score tests in heteroscedastic *t* regression models. 2000. 31p. (RT-MAE-2000-7)

2000-8 - **PEREIRA, C.A.B., NAKANO, F., STERN, J.M.** Actuarial Analysis via Branching Processes. 2000. 7p.(RT-MAE-2000-8)

2000-9 - **VIANA, M.A.G., PEREIRA, C.A.DE B.** Statistical Assessment of Jointly Observed Screening Tests. 2000 13p. (RT-MAE-200-9)

2000-10 - **MADRUGA, M.R., ESTEVES, L.G., WECHSLER, S.** On the Bayesianity of Pereira-Stern Tests. 2000. 9p. (RT-MAE-2000-10)

2000-11 - **PEREIRA, C.A., STERN, J.M.** *Full Bayesian Significance Test for Coefficients of Variation. 2000. 7p. (RT-MAE-2000-11)*

2000-12 - **BOLFARINE, H, CABRAL, C.R.B., PAULA, G.A.** Distance Tests Under Nonregular Conditions: Applications to the Comparative Calibration Model. 2000. 15p. (RT-MAE-2000-12)

2000-13 - **FERRARI, S.L.P., URIBE-OPAZO, M.** Corrected Likelihood Ratio Tests in a Class of Symmetric Linear Regression Models. 2000. 18p. (RT-MAE-2000-13)

2000-14 - **SVETLIZA, C.F., PAULA, G.A.** On Diagnostics in Log-Linear Negative Binominal Models. 2000. 17p. (RT-MAE-2000-14)

2000-15 - **FONTES, L.R.G., ISOPI, M., NEWMAN, C.M.** Random walks with strongly inhomogeneous rates and singular diffusions: convergence, localization and aging in one dimension. 2000. 22p. (RT-MAE-2000-15)

2000-16 - **SAHU, S.K., DEY, D.K., BRANCO, M.D.** A New Class of Multivariate Skew Distributions with Applications to Bayesian Regression Models. 2000. 28p. (RT-MAE-2000-16)

2000-17 - **PEREIRA, C.A.B., STERN, J.M.** Model Selection: Full Bayesian Approach. 2000. 13p. (RT-MAE-2000-17)

2000-18 - **CORDEIRO, G.M., BOTTER, D.A.** Second-Order biases of maximum likelihood estimates in overdispersed generalized linear models. 2000. 16p. (RT-MAE-2000-18)

2000-19 - **GONZALEZ-LOPES, V.A., TANAKA, N.I.** Maximal Association, Fréchet Bounds and Homotopy in Prediction. 2000. 21p. (RT-MAE-2000-19)

2000-20 - **FERRARI, S.L.P., CRIBARI-NETO, F.** Corrected modified profile likelihood heteroskedasticity tests. 2000. 11p. (RT-MAE-2000-20)

2000-21 - **ESTEVES, L.G., WECHSLER, S., IGLESIAS, P.L.** De Finettian theorems for planar uniform models. 2000. 40p. (RT-MAE-2000-21)

2000-22 - **BUENO, V.C.** Modeling a coherent system at critical level. 2000. 9p. (RT-MAE-2000-22)

2000-23 - **ANDRÉ, C.D.S., ELIAN, S.N.**, Comparação de Estimadores do Parâmetro de Escala da Distribuição dos Erros na Regressão L_1. 2000. 28p. (RT-MAE-2000-23)

The complete list of "Relatórios do Departamento de Estatística", IME-USP, will be sent upon request.

Departamento de Estatística
IME-USP
Caixa Postal 66.281
05315-970 - São Paulo, Brasil