A Bayesian approach to the analysis of quantal bioassay studies using nonparametric mixture models

Kassandra Fronczyk[1,*] **and Athanasios Kottas**[2,**]

[1]Department of Statistics, Rice University, Houston, Texas 77030, U.S.A.

[2]Department of Applied Mathematics and Statistics, University of California, Santa Cruz, California 95064, U.S.A.

[*]*email:* kassandra.fronczyk@gmail.com

[**]*email:* thanos@ams.ucsc.edu

SUMMARY. We develop a Bayesian nonparametric mixture modeling framework for quantal bioassay settings. The approach is built upon modeling dose-dependent response distributions. We adopt a structured nonparametric prior mixture model, which induces a monotonicity restriction for the dose-response curve. Particular emphasis is placed on the key risk assessment goal of calibration for the dose level that corresponds to a specified response. The proposed methodology yields flexible inference for the dose-response relationship as well as for other inferential objectives, as illustrated with two data sets from the literature.

KEY WORDS: Calibration; Cytogenetic dosimetry; Dependent Dirichlet process; Dose-response curve; Markov chain Monte Carlo; Nonparametric mixture models.

1. Introduction

The quantal bioassay dose-response setting has been valuable in many fields, from ecology to medicine. Though quantitative measurement of a response is preferred, there are several experiments for which the response can only be recorded on a binary scale, either occurring or not occurring. The dose-response relationship is based on observed data from experimental animal, human clinical, or cell studies. In these types of settings, as the level of exposure increases, the responses generally become more severe. However, within a population, a wide range of responses may be encountered as some individuals are more susceptible and others resistant.

Regarding the quantal response data structure, at each of a typically small number of dose levels, a number of subjects are exposed to the substance and the subset that have the response of interest is recorded. The dose-response curve, denoted by $D(x)$, is defined as the probability of positive response as a function of dose x. One of the standard assumptions in the analysis of quantal bioassay problems is that $D(x)$ is a non-decreasing function and can therefore be modeled as a distribution function. Under this assumption, the distribution corresponding to $D(x)$ is referred to as the tolerance distribution. In parametric modeling, $D(x)$ is assumed to be a member of a parametric family of distribution functions. Standard parametric models, such as probit or logit models, are simple to implement but do not have the ability to capture complex curves, including skewness and multimodality.

To extend the inferential scope of standard models, finite mixtures for the tolerance distribution have been considered in the literature. Lwin and Martin (1989) present a location-scale mixture model, which is developed based on the idea of a finite number of unobserved, underlying subpopulations. Geweke and Keane (1999) propose mixtures of probit models to more accurately describe the shape of the dose-response curve,

while Basu and Mukhopadyay (2000) use a finite scale mixture of normal distribution functions. In addition to theoretical advantages, Bayesian nonparametric mixture priors, as developed in this paper, offer practical benefits relative to finite parametric mixtures with respect to Markov chain Monte Carlo (MCMC) posterior simulation as well as prior specification.

The quantal bioassay setting corresponds to one of the earliest applications for Bayesian nonparametrics. One of the more common approaches involves assigning a Dirichlet process (DP) prior (Ferguson, 1973) to the distribution function $D(x)$ (e.g., Antoniak, 1974; Bhattacharya, 1981; Disch, 1981; Kuo, 1983, 1988; Gelfand and Kuo, 1991; Mukhopadhyay, 2000). Also along these lines is the work in Muliere and Walker (1997) where Pólya tree priors are used to determine the maximum tolerated dose. A monotone nonparametric regression framework is presented in Bornkamp and Ickstadt (2009), where the monotone function is modeled as a mixture of shifted and scaled two-sided power distribution functions.

An alternative approach is to model the binary responses with latent continuous variables, and assume mixtures of probit or logit functions for the dose-response curve. MacEachern (1998) assumes a DP prior directly on the distribution of the latent variables. Presented in terms of multivariate probit regression, Jara et al. (2007) mix on the intercept of the linear mean and the covariance matrix of the distribution of the subject specific latent variable distribution. Casanova et al. (2010) place a scale normal DP mixture prior on the distribution of the latent variables.

We also represent the model through latent continuous responses, but build the inference for the dose-response curve from nonparametric mixture modeling for dose-dependent response distributions. We adopt a structured dependent DP prior for the collection of dose-dependent mixing distributions, which begets a monotonicity

restriction and produces smooth realizations of the dose-response curve. We place particular emphasis on inference for calibration, where we seek to estimate the dose level associated with a given vector of responses. This objective is especially important in cytogenetic dosimetry, a particular area of dose-response modeling concerned with the relationship between exposure to radiation and some measure of genetic aberration. In these studies, only a portion of the dose-response curve is observed, rendering extrapolation beyond observed dose levels a primary inferential target.

The outline of the paper is as follows. We develop the model in Section 2, including the hierarchical formulation with latent continuous responses, and a method for prior specification. The approach to posterior simulation and inference for risk assessment is provided in Section 3. In Section 4, we illustrate the modeling approach with a standard data set from the literature as well as with the cytogenetic dosimetry application. Finally, Section 5 concludes with discussion of possible extensions of the methodology.

2. Methods

2.1 The Modeling Approach

The starting point of our methodology is nonparametric mixture modeling for the collection of response distributions, which are indexed by dose level $x \in \mathcal{X}$, where, typically, $\mathcal{X} \subset \mathbb{R}$, as dose is commonly recorded on a logarithmic scale. With replicated binary responses at a number of observed dose values, the natural mixture structure involves mixing of Bernoulli distributions,

$$f(y \mid G_x) = \int \text{Bern}\left(y \mid \Phi(\theta)\right) \mathrm{d}G_x(\theta), \quad x \in \mathcal{X} \tag{1}$$

where $\Phi(\cdot)$ is the standard normal distribution function, and thus $\theta \in \mathbb{R}$. Given

the mixture setting, the choice of the link function for the kernel probability is driven mainly by convenience in model implementation where the probit link offers advantages. We aim to develop flexible nonparametric modeling for the response distributions, $\{f(y \mid G_x) : x \in \mathcal{X}\}$, and for the implied dose-response relationship, $\Pr(y = 1 \mid G_x)$, $x \in \mathcal{X}$, retaining however the traditional assumption of monotonicity for the dose-response curve. To this end, the model needs to be completed with an appropriate nonparametric prior for the collection of mixing distributions, $G_{\mathcal{X}} = \{G_x : x \in \mathcal{X}\}$.

Given the nature of quantal bioassay studies, we expect response distributions associated with nearby dose levels to be more similar than those which are far apart. Under our mixture model setting, this situation gives rise to the need for a prior which relates the mixing distributions G_x across dose level x to varying degrees. A powerful option for this modeling problem is the dependent Dirichlet process (DDP) prior (MacEachern, 2000). The DDP prior is motivated by the (almost sure) discrete representation of DP realizations (Sethuraman, 1994), where, in full generality, both the stick-breaking weights and the atoms may evolve with x. Because the general DDP prior is complicated to implement and requires large data sets to sufficiently learn about its hyperparameters, simpler versions are typically employed in applications. Under minimal conditions, Barrientos et al. (2012) have established full support (according to weak convergence) for both of the simplified DDP prior specifications discussed next.

In particular, a "single-p" DDP prior involves a countable mixture of realizations from a stochastic process over $\mathcal{X}$, with weights matching those from the standard DP, that is, $G_{\mathcal{X}} = \sum_{l=1}^{\infty} \omega_l \delta_{\eta_{l\mathcal{X}}}$. Here, the $\eta_{l\mathcal{X}} = \{\eta_l(x) : x \in \mathcal{X}\}$ are i.i.d. realizations from a base stochastic process, $G_{0\mathcal{X}}$, over $\mathcal{X}$, and the weights arise from stick-breaking: $\omega_1 = \zeta_1$, $\omega_l = \zeta_l \prod_{r=1}^{l-1}(1 - \zeta_r)$ for $l \geq 2$, with the ζ_l i.i.d. from a Beta$(1, \alpha)$ distribution (independently of the $\eta_{l\mathcal{X}}$). Under a single-p DDP prior for the mixing distributions

in (1), the dose-response curve, $\Pr(y = 1 \mid G_x) = \int \Phi(\theta) \mathrm{d}G_x(\theta) = \sum_{l=1}^{\infty} \omega_l \Phi(\eta_l(x))$, $x \in \mathcal{X}$. This model does not enforce monotonic dose-response relationships, although an increasing trend can be incorporated in prior expectation. This structure can be achieved through a Gaussian process for $G_{0\mathcal{X}}$ with constant variance, σ^2, and linear mean function, $\beta_0 + \beta_1 x$, with $\beta_1 > 0$. Then, the $\mathrm{N}(\beta_0 + \beta_1 x, \sigma^2)$ distribution for G_{0x}, induced by the Gaussian process $G_{0\mathcal{X}}$ at $x \in \mathcal{X}$, is stochastically ordered, and thus $\mathrm{E}(\Pr(y = 1 \mid G_x)) = \int \Phi(\theta) \mathrm{d}G_{0x}(\theta)$ is increasing in x being the expectation of the increasing function $\Phi(\theta)$ with respect to G_{0x}. (Depending on the context, we use $\mathrm{N}(m, s^2)$ for either the normal distribution or density with mean m and variance s^2.)

Contrarily, we may consider a DDP prior structure where only the weights evolve with x, that is, $G_{\mathcal{X}} = \sum_{l=1}^{\infty} \omega_{l\mathcal{X}} \delta_{\theta_l}$, with the θ_l i.i.d. from a base distribution G_0 on $\mathbb{R}$, independently of the stochastic mechanism that generates the $\omega_{l\mathcal{X}} = \{\omega_l(x) : x \in \mathcal{X}\}$. This "single-$\theta$" DDP formulation presents a formidable complication with regard to the main inferential objectives for bioassay experiments. A monotonically increasing dose-response relationship, at least in prior expectation, is imperative in terms of prediction at unobserved dose levels to anchor the inference with an increasing trend. In the case of the single-θ DDP prior, there is no means to force such a trend for the dose-response curve, $\Pr(y = 1 \mid G_x) = \sum_{l=1}^{\infty} \omega_l(x) \Phi(\theta_l)$. More specifically, using the monotone convergence theorem and letting $C = \int \Phi(\theta) \mathrm{d}G_0(\theta)$ (where a finite mean is assumed for G_0), we obtain $\mathrm{E}(\Pr(y = 1 \mid G_x)) = \sum_{l=1}^{\infty} \mathrm{E}(\Phi(\theta_l)) \mathrm{E}(\omega_l(x)) = C \sum_{l=1}^{\infty} \mathrm{E}(\omega_l(x)) = C$, for any $x \in \mathcal{X}$. Thus, the prior expectation of the dose-response curve is constant in x, rendering interpolation and extrapolation inference practically useless.

Hence, the single-p DDP prior emerges as the preferred choice for dose-response inference built from modeling dependent response distributions. The general version of the single-p DDP prior can be used for settings where one seeks the extra flexibility

of non-monotonic dose-response functions with the increasing trend only in prior expectation. We have studied such DDP mixture modeling for developmental toxicity experiments, which involve clustered categorical responses and multiple dose-response curves for distinct endpoints (Fronczyk and Kottas, 2013; Kottas and Fronczyk, 2013). For these more involved experiments, the capacity of the DDP mixture model to uncover non-monotonic dose-response relationships is a practically relevant feature. However, for the simpler quantal bioassay setting, we seek a more structured nonparametric prior to incorporate the standard monotonicity assumption for the dose-response curve with prior probability one rather than only in prior expectation.

To this end, we explore the linear-DDP prior formulation, which simplifies the single-p DDP stochastic process realizations $\{\eta_l(x) : x \in \mathcal{X}\}$ to linear functions of x, that is, $\eta_l(x) = \gamma_{0l} + \gamma_{1l}x$. Hence, the base stochastic process $G_{0\mathcal{X}}$ is replaced with a distribution G_0 that generates the component specific intercept and slope parameters. In particular, we assume the γ_{0l} are i.i.d. from $G_0^{(0)}$, and independently, the γ_{1l} are i.i.d. from $G_0^{(1)}$. We refer to DeIorio et al. (2009) for survival regression with linear-DDP priors, and DeIorio et al. (2004) for the closely related ANOVA-DDP model.

The essential observation from the construction of the linear-DDP prior is the correspondence of $\{G_x : x \in \mathcal{X}\}$ and $G = \sum_{l=1}^{\infty} \omega_l \delta_{(\gamma_{0l},\gamma_{1l})}$, where $G \sim \mathrm{DP}(\alpha, G_0)$ with $G_0 = G_0^{(0)} G_0^{(1)}$. In particular, the atoms $(\gamma_{0l}, \gamma_{1l})$, sampled originally from G_0, and the weights, ω_l, in the stick-breaking construction of G induce the atoms $\eta_l(x) = \gamma_{0l} + \gamma_{1l}x$ in conjunction with the same weights, ω_l, in the stick-breaking construction of G_x, for all x. Hence, the linear-DDP mixture model, $f(y \mid G_x) = \int \mathrm{Bern}(y \mid \Phi(\theta))\mathrm{d}G_x(\theta)$, where $G_x = \sum_{l=1}^{\infty} \omega_l \delta_{(\gamma_{0l}+\gamma_{1l}x)}$, can be equivalently formulated as a DP mixture model

$$f(y \mid G, x) = \int \mathrm{Bern}(y \mid \Phi(\gamma_0 + \gamma_1 x))\mathrm{d}G(\gamma_0, \gamma_1); \quad G \sim \mathrm{DP}(\alpha, G_0 = G_0^{(0)} G_0^{(1)}).$$

The corresponding dose-response curve is now defined by

$$\Pr(y = 1 \mid G, x) = \int \Phi(\gamma_0 + \gamma_1 x) \mathrm{d}G(\gamma_0, \gamma_1) = \sum\nolimits_{l=1}^{\infty} \omega_l \Phi(\gamma_{0l} + \gamma_{1l} x), \quad x \in \mathcal{X}. \quad (2)$$

Hence, under the more structured linear-DDP mixture model, the restriction $\gamma_1 > 0$, implemented with a distribution $G_0^{(1)}$ supported by $\mathbb{R}^+$, yields a sufficient condition for dose-response curve realizations to be increasing with prior probability one.

For most bioassay experiments, it is natural to view the observed binary response as a proxy for a latent continuous response variable. This procedure offers a tractable way to estimate the parameters of the proposed DP mixture model. In particular, with the restriction $\gamma_1 > 0$, we can reparameterize the Bernoulli kernel, and, consequently, the underlying latent response distribution, to a location-scale mixture of normal distributions. Specifically, let $\mu = -\gamma_0/\gamma_1 \in \mathbb{R}$ and $\tau = 1/\gamma_1 > 0$, and denote by z the latent $\mathbb{R}$-valued response associated with the binary outcome y. Then, we obtain

$$\begin{aligned}
\Pr(y = 1 \mid G, x) &= \int \Phi(\gamma_0 + \gamma_1 x) \mathrm{d}G(\gamma_0, \gamma_1) \\
&= \int \Phi\left((-\mu/\tau) + (1/\tau)x\right) \mathrm{d}G^*(\mu, \tau^2) = \int \left\{ \int_{-\infty}^{x} \mathrm{N}(z \mid \mu, \tau^2) \mathrm{d}z \right\} \mathrm{d}G^*(\mu, \tau^2) \\
&= \int_{-\infty}^{x} \left\{ \int \mathrm{N}(z \mid \mu, \tau^2) \mathrm{d}G^*(\mu, \tau^2) \right\} \mathrm{d}z \;=\; \Pr(z \leq x \mid G^*)
\end{aligned}$$

where the final probability arises under a location-scale normal DP mixture for the latent tolerance distribution, $\int \mathrm{N}(z \mid \mu, \tau^2) \mathrm{d}G^*(\mu, \tau^2)$, where $G^* \sim \mathrm{DP}(\alpha, G_0^*)$ with $G_0^* \equiv G_0^*(\mu, \tau^2)$ an appropriate distribution on $\mathbb{R} \times \mathbb{R}^+$. Because the (μ, τ^2) parameterization expedites prior specification, we work with the equivalent location-scale normal DP mixture formulation (suppressing the G^* notation in the following).

2.2 Hierarchical Model Formulation

The general structure for the observations comprises data $= \{(x_i, y_{ij}) : i = 1, ..., N; j = 1, ..., n_i\}$, where y_{ij} are the binary responses from the n_i subjects at dose level x_i. Moreover, we denote by z_{ij} the latent (unobserved) continuous response corresponding to y_{ij}. Inclusion of the latent responses facilitates posterior simulation; see Section 3. Hereinafter, we use the notation $\mathbf{1}_{ij}(y_{ij})$ to indicate the relationship between y_{ij} and z_{ij}, where $\mathbf{1}_{ij}(y_{ij}) = 1$ when $z_{ij} \leq x_i$, and $\mathbf{1}_{ij}(y_{ij}) = 0$ otherwise.

Using the location-scale parameterization for the normal DP mixture tolerance distribution developed in Section 2.1, the mixture model for the data is given by

$$
\begin{aligned}
y_{ij} \mid z_{ij} &\overset{ind.}{\sim} \mathbf{1}_{ij}(y_{ij}), \;\; i = 1, \ldots, N; \;\; j = 1, \ldots, n_i \\
z_{ij} \mid G &\overset{ind.}{\sim} \int \mathrm{N}(z_{ij} \mid \mu, \tau^2) \mathrm{d}G(\mu, \tau^2), \;\; i = 1, \ldots, N; \;\; j = 1, \ldots, n_i
\end{aligned}
$$

with $G \mid \alpha, \boldsymbol{\psi} \sim \mathrm{DP}(\alpha, G_0(\mu, \tau^2 \mid \boldsymbol{\psi}))$ and $G_0(\mu, \tau^2 \mid \boldsymbol{\psi})$ defined through independent $\mathrm{N}(\mu|\beta_0, \sigma_0^2)$ and inv-gamma$(\tau^2|c, \delta_0)$ components, where the inverse gamma distribution has mean $\delta_0/(c-1)$ provided $c > 1$. Here, c is fixed and $\boldsymbol{\psi} = (\beta_0, \sigma_0^2, \delta_0)$ denotes the parameters of the DP base distribution which are assigned hyperpriors.

As a computational tool, we use a DP truncation approximation replacing G with $G^L = \sum_{l=1}^{L} p_l \delta_{(\mu_l, \tau_l^2)}$. Truncation from the outset provides closed form full conditionals for posterior simulation via Gibbs sampling and a straightforward means for obtaining inference for all objectives. Here, the (μ_l, τ_l^2), given $\boldsymbol{\psi}$, are i.i.d. G_0, and the weights p_l arise from a truncated version of the stick-breaking construction: $p_1 = V_1$, $p_l = V_l \prod_{r=1}^{l-1}(1 - V_r)$, $l = 2, \ldots, L-1$, and $p_L = 1 - \sum_{l=1}^{L-1} p_l$, with the V_l i.i.d., given α, from Beta$(1, \alpha)$. The truncation level can be chosen using distributional properties for the tail probability of the stick-breaking weights, $U_L = \sum_{l=L+1}^{\infty} \omega_l$. More specifically,

$\mathrm{E}(U_L \mid \alpha) = \{\alpha/(\alpha+1)\}^L$ and $\mathrm{E}(U_L^2 \mid \alpha) = \{\alpha/(\alpha+2)\}^L$ (Ishwaran and Zarepour, 2000). These expressions can be averaged over the prior for α to estimate $\mathrm{E}(U_L)$ and $\mathrm{Var}(U_L)$. Given a tolerance level for the approximation, the former expression yields the corresponding value L. For both data examples of Section 4, we used a gamma$(2,1)$ prior for α. Consequently, we set the truncation level to $L = 50$, which, after averaging over the prior for α, results in $\mathrm{E}(U_{50}) \approx 0.000045$, and $\mathrm{Var}(U_{50}) \approx 0.00000038$.

To represent the mixture component with which each data point y_{ij} is associated, configuration variables w_{ij} are introduced. Then, the hierarchical model for the data, augmented with the latent continuous responses, can be written as follows:

$$\begin{aligned} y_{ij} \mid z_{ij} &\overset{ind.}{\sim} \mathbf{1}_{ij}(y_{ij}), \quad i = 1, \ldots, N;\ j = 1, \ldots, n_i \\ z_{ij} \mid \boldsymbol{\mu}, \boldsymbol{\tau}^2, \boldsymbol{w} &\overset{ind.}{\sim} \mathrm{N}(\mu_{w_{ij}}, \tau^2_{w_{ij}}), \quad i = 1, \ldots, N;\ j = 1, \ldots, n_i \\ w_{ij} \mid \boldsymbol{p} &\overset{ind.}{\sim} \textstyle\sum_{l=1}^{L} p_l \delta_l, \quad i = 1, \ldots, N;\ j = 1, \ldots, n_i \\ (\mu_l, \tau_l^2) \mid \boldsymbol{\psi} &\overset{ind.}{\sim} G_0, \quad l = 1, \ldots, L, \end{aligned} \tag{3}$$

where the prior distribution for $\boldsymbol{p} = (p_1, ..., p_L)$ is given by $f(\boldsymbol{p} \mid \alpha) = \alpha^{L-1} p_L^{\alpha-1}(1-p_1)^{-1}(1-(p_1+p_2))^{-1} \times \cdots \times (1 - \sum_{l=1}^{L-2} p_l)^{-1}$. Here, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_L)$, $\boldsymbol{\tau}^2 = (\tau_1^2, \ldots, \tau_L^2)$, and $\boldsymbol{w} = \{w_{ij} : i = 1, \ldots, N;\ j = 1, \ldots, n_i\}$. In addition to the gamma prior for α, we place normal $\mathrm{N}(m_{\beta_0}, s^2_{\beta_0})$ and inv-gamma$(A_{\sigma_0^2}, B_{\sigma_0^2})$ priors on β_0 and σ_0^2, respectively, and an exponential prior on δ_0 with mean B_{δ_0}.

2.3 Prior Specification

Regarding the DP base distribution parameters, $\boldsymbol{\psi} = (\beta_0, \sigma_0^2, \delta_0)$, empirical evidence suggests robustness to their hyperprior choice. In general, we recommend setting the shape parameter, c, of the inverse gamma distribution for the τ_l^2, and of the inverse gamma prior for σ_0^2 to values that yield relatively dispersed distributions, but with

finite variance. Then, working with the expressions for the marginal mean, $\mathrm{E}(z)$, and variance, $\mathrm{Var}(z)$, of the latent responses under a single component of the mixture model, we are able to define the hyperpriors through only the range of the dose levels. In particular, $\mathrm{E}(z) = \mathrm{E}\{\mathrm{E}(z|\mu, \tau^2)\} = \mathrm{E}(\mu) = \mathrm{E}\{\mathrm{E}(\mu|\beta_0, \sigma_0^2)\} = \mathrm{E}(\beta_0) = m_{\beta_0}$, and thus m_{β_0} can be specified by the midrange of the doses. Analogously, $\mathrm{Var}(z) = \mathrm{Var}(\mu) + \mathrm{E}(\tau^2) = \mathrm{Var}(\beta_0) + \mathrm{E}(\sigma_0^2) + (c-1)^{-1}\mathrm{E}(\delta_0) = s^2_{\beta_0} + (A_{\sigma_0^2} - 1)^{-1}B_{\sigma_0^2} + (c-1)^{-1}B_{\delta_0}$. Therefore, with c and $A_{\sigma_0^2}$ specified, the variance components, $s^2_{\beta_0}$, $B_{\sigma_0^2}$ and B_{δ_0}, can be determined through a proxy for the marginal variance obtained by dividing the range of the dose levels by 4 (or 6) and squaring.

The DP precision parameter α controls the number of distinct components, n^*, in mixture model (3), that is, the number of distinct w_{ij} associated with the $n' = \sum_{i=1}^{N} n_i$ binary outcomes. In particular, for moderately large n', $\mathrm{E}(n^* \mid \alpha) = \alpha \log\{(\alpha + n')/\alpha\}$, and such expressions can be used to guide prior choice for α. For the data examples of Section 4.1 and 4.2, the gamma$(2, 1)$ prior for α implies $\mathrm{E}(n^*) \approx 10$ and 16, respectively.

3. Posterior Inference

3.1 MCMC Posterior Simulation

We use blocked Gibbs sampling (Ishwaran and James, 2001) for simulation from the posterior distribution of the DP mixture model in (3). Posterior samples for the model parameters are used in inference for risk assessment as discussed in Section 3.2.

Each latent response z_{ij} is imputed based on the value of y_{ij}. If $y_{ij} = 0$ ($y_{ij} = 1$), z_{ij} is drawn from a $\mathrm{N}(\mu_{w_{ij}}, \tau^2_{w_{ij}})$ distribution truncated below (above) at x_i.

We denote the n^* distinct values of $\boldsymbol{w}$ by $w_1^*, \ldots, w_{n^*}^*$, and let $M_k^* = |\{(i,j) : w_{ij} = w_k^*\}|$, for $k = 1, \ldots, n^*$, and $M_l = |\{(i,j) : w_{ij} = l\}|$, for $l = 1, \ldots, L$. Then, if l does not correspond to a distinct component, the (μ_l, τ_l^2) are sampled, given $\boldsymbol{\psi}$, from G_0. For

the active components, that is, for $l \in \{w_k^* : k = 1, \ldots, n^*\}$, $\mu_{w_k^*}$ arises from a normal distribution with mean $(\sigma_0^{-2}\beta_0 + \tau_{w_k^*}^{-2} \sum \sum_{\{(i,j):w_{ij}=w_k^*\}} z_{ij})/(\tau_{w_k^*}^{-2} M_k^* + \sigma_0^{-2})$ and variance $(\tau_{w_k^*}^{-2} M_k^* + \sigma_0^{-2})^{-1}$. Similarly, $\tau_{w_k^*}^2$ is drawn from an inverse gamma distribution with shape parameter $c + 0.5 M_k^*$ and scale parameter $\delta_0 + 0.5 \sum \sum_{\{(i,j):w_{ij}=w_k^*\}} (z_{ij} - \mu_{w_k^*})^2$.

Each subject-specific configuration variable w_{ij} is drawn from a discrete distribution with probabilities proportional to $p_l \mathrm{N}(z_{ij} \mid \mu_l, \tau_l^2)$, for $l = 1, \ldots, L$. The updates for the vector of weights, $\boldsymbol{p}$, and the precision parameter, α, are the same with a generic DP mixture model (e.g., Ishwaran and Zarepour, 2000).

The conditionally conjugate priors used for the parameters of G_0 lead to standard updates for the components of $\boldsymbol{\psi}$. Specifically, the posterior full conditional for β_0 is normal with mean $(s_{\beta_0}^{-2} m_{\beta_0} + \sigma_0^{-2} \sum_{l=1}^{L} \mu_l)/(\sigma_0^{-2} L + s_{\beta_0}^{-2})$ and variance $(\sigma_0^{-2} L + s_{\beta_0}^{-2})^{-1}$, and for σ_0^2 it is given by an inverse gamma distribution with shape parameter $0.5L + A_{\sigma_0^2}$ and scale parameter $B_{\sigma_0^2} + 0.5 \sum_{l=1}^{L} (\mu_l - \beta_0)^2$. Finally, δ_0 has a gamma posterior full conditional with shape parameter $Lc + 1$ and rate parameter $B_{\delta_0}^{-1} + \sum_{l=1}^{L} \tau_l^{-2}$.

3.2 Inference for Risk Assessment

The most important inference under the quantal bioassay setting is for the dose-response relationship. Expressed in terms of the tolerance distribution, and under the DP truncation approximation, the dose-response curve in (2) becomes

$$D(x) \equiv \Pr(y = 1 \mid G, x) = \Pr(z \leq x \mid G) = \sum_{l=1}^{L} p_l \Phi\left((x - \mu_l)/\tau_l\right).$$

Section 2.1 elucidates the monotonicity property for the dose-response curve in terms of the $\gamma_1 > 0$ restriction, which carries over to the tolerance distribution formulation.

Another inferential objective is inversion, where interest lies in estimation of the

dose-level, x_q, corresponding to a specified probability, q. Posterior draws for x_q are obtained by numerically inverting the posterior realization of the dose-response curve at each iteration of the MCMC algorithm.

An important feature of the linear-DDP mixture modeling framework is that it lends itself to inference for calibration, which is a key risk assessment inferential objective for certain types of experiments. Here, we consider a specified vector of responses, $\boldsymbol{y}_0 = \{y_{0j} : j = 1, ..., n_0\}$ and seek to estimate the dose level, x_0, which is associated with this new vector of responses. This inference can be obtained by augmenting the hierarchical model with the components associated with $\boldsymbol{y}_0$ and expanding the parameter vector to include the unknown dose x_0. Because the DDP prior is defined over the uncountable space $\mathcal{X}$, it induces a proper hierarchical model formulation for any x_0. Specifically, the full hierarchical model, $p(\{y_{ij}\} \mid \{z_{ij}\}, x_0)p(\{z_{ij}\} \mid \{w_{ij}\}, \boldsymbol{\mu}, \boldsymbol{\tau}^2)p(\{w_{ij}\} \mid \boldsymbol{p})$, for calibrating dose x_0 is given by:

$$\prod_{i=0}^{N}\prod_{j=1}^{n_i}\left\{\mathbf{1}_{ij}(y_{ij})\mathrm{N}(z_{ij} \mid \mu_{w_{ij}}, \tau^2_{w_{ij}})\sum_{l=1}^{L} p_l\delta_l(w_{ij})\right\}$$

where the prior for the (μ_l, τ_l^2) and for $\boldsymbol{p}$, and the hyperpriors for α and $\boldsymbol{\psi}$, are equivalent to model (3). In Section 4, we experiment with both normal and uniform priors, $p(x_0)$, for x_0, obtaining empirical evidence that a plausible prior range for the calibrated dose level suffices for robust posterior inference.

Regarding posterior simulation, conditional on x_0, the main MCMC updates are as in Section 3.1 now extended to $i = 0, 1, \ldots, N$. The full conditional for x_0 is proportional to $p(x_0)\prod_{j=1}^{n_0}\mathbf{1}_{0j}(y_{0j})$, where $\mathbf{1}_{0j}(y_{0j}) = 1$ (0) if $z_{0j} \leq x_0$ ($z_{0j} > x_0$). Hence, draws from the full conditional of x_0 arise from a truncated version of its prior distribution, truncated below at $\max\{z_{0j} : y_{0j} = 1\}$ and above at $\min\{z_{0j} : y_{0j} = 0\}$.

4. Data Illustrations

The performance of the linear-DDP mixture model was evaluated using two synthetic data sets generated from tolerance distributions that exhibit skewness and multimodality. Inference results were also contrasted with the single-p DDP prior setting, discussed in Section 2.1, which supports increasing dose-response curves only in prior expectation. In both cases, the linear-DDP model recovers successfully the true dose-response curve with narrower uncertainty bands than the general DDP prior model. Details from this simulation study are provided in Web Appendix A.

Here, we illustrate the methodology with two data sets, one studied throughout the literature and the other an application to cytogenetic dosimetry.

4.1 Trypanosome Data

Ashford and Walker (1972) first analyzed data reporting the death rate of protozoan trypanosome found in Table 1. Trypanosomes are parasites which do not grow readily in artificial culture (when they do grow, they change from the blood form to the insect form) and no reliable quantitative genetic characters have been found nor have accurate methods been available to assess characters such as drug resistance. Here, we look at results from the trypanosome sensitivity to the log concentrations of pure neutral acriflavine (Walker, 1966).

Following the prior specification approach discussed in Section 2.3, we implemented model (3) setting $c = 5$ and placing a $\text{N}(5.1, 0.2^2)$ prior on β_0, an inv-gamma$(4, 0.07)$ prior on σ_0^2, and an exponential prior with mean 0.2 on δ_0.

The posterior mean and 90% uncertainty bands for the dose-response curve are shown in the left panel of Figure 1. The posterior mean estimate follows closely the path of the observed data, and the interval bands include the majority of the points.

The extremes of the observed proportions are at 0 and 1, which are known to be burdensome for most parametric models. In general, the linear-DDP mixture model captures well the non-standard, bimodal shape for the tolerance distribution.

The middle panel of Figure 1 includes the posterior estimates for dose levels, x_q, for inversion probabilities $q = 0.05, 0.15, 0.25$, and 0.5. The spread of the densities depends on the width of the probability intervals around the dose-response curve at the given response probability. The posterior range for $x_{0.25}$ is comparable to that obtained by Mukhopadhyay (2000), using a DP prior for the dose-response function.

Finally, we consider inference for the calibrated dose level, x_0, which corresponds to a new response vector comprising 26 positive responses from 52 subjects, that is, $n_0 = 52$ and $\sum_{j=1}^{n_0} y_{0j} = 26$, in the notation of Section 3.2. This new response vector resembles the observed counts at dose 5.1, thus offering a useful setting for study of sensitivity to the prior specification for x_0. The right panel of Figure 1 plots the posterior density for x_0 under three different priors: a normal prior with mean 5.1 and standard deviation 0.07, a uniform prior on $(4.9, 5.25)$, and a uniform prior over the entire dose range. The general uniform prior results in the largest spread and the spread decreases as the uniform support shrinks or switches to the normal distribution. The specific normal prior leads to a relatively small amount of learning in the resulting posterior density for the calibrated dose level. However, it is encouraging that the normal prior and the uniform prior on the relatively wide range $(4.9, 5.25)$ yield similar posterior densities.

4.2 Cytogenetic Dosimetry Application

Cytogenetic dosimetry is a biological tool for dose assessment in cases of radiological accidents and suspected overexposures. The main focus of these studies is to determine

the relationship between the dosage of exposure to radiation and some measure of genetic aberration. The typical experiment considers samples of cell cultures exposed to a range of levels of an agent with some measure of cell disability recorded as the response. The two main inferential objectives are prediction of response at unobserved exposure levels, and inference for unknown dose levels given observed responses.

For an illustration, we consider part of a larger data set where blood samples from individuals were exposed to ^{60}Co radiation at doses of 0, 20, 50, 100, 200, 300, 400, and 500 cGy (centograms). The resulting cultures were tested for binucleated cells with a recorded number of micronuclei (MN). The full data set, found in Madruga et al. (1996), groups the outcomes into none, one, or two or more MN; see Table 2 for the subset we work with here from healthy, older subjects. This particular ordinal classification is used because when many MN are present, they can be difficult to count exactly. We collapse to two classification groups, no MN and one or more MN, in order to apply the linear-DDP Bernoulli mixture model. In particular, model (3) is fitted setting $c = 5$ and using a $\mathrm{N}(200, 85^2)$ prior for β_0, an inv-gamma$(4, 500)$ prior for σ_0^2, and an exponential prior with mean 500 for δ_0.

Plotted in the top panels of Figure 2 are posterior estimates of the dose-response curve, and the inversion dose levels for $q = 0.05, 0.15, 0.25$ and 0.3. The 90% probability bands of the dose-response curve are much tighter than those found in the trypanosome data, as here the number of subjects per dose level is roughly 20 times larger. The data is only observed to about 40% of the full curve, and therefore the intervals get much wider beyond the largest experimental dose level. The inversion inference is interesting in that at the smaller levels of q (0.05 and 0.15), the estimates of the corresponding dose level have relatively narrow densities, whereas the larger values (0.25 and 0.3) result in densities with heavier tails and substantially larger spread. This coincides

with the uncertainty bands around the dose-response curve in these areas.

To illustrate inference for calibration, we consider two vectors of new responses which are equivalent to the observed vectors at 400 cGy and 500 cGy. In each case, we used a relatively dispersed normal prior for x_0 centered at the corresponding observed dose level with variance 45^2. Figure 2 provides results from the model fitted to the full data set in Table 2, but also to the reduced data with the respective observations at 400 cGy and 500 cGy excluded. In the former case (Figure 2, bottom left panel), the calibrated dose level posterior density captures fairly well the value of 400 cGy, and interestingly, with a relatively small increase in uncertainty when the corresponding data vector is not included in the model fit. For the 500 cGy dose level, excluding the respective observations yields a challenging extrapolation test for the model with data available up to dose 400 cGy and with less than 35% of the dose-response curve observed. The value of 500 cGy is contained in the support of the resulting calibrated dose distribution, but the corresponding density (Figure 2, bottom right panel) shows a more noticeable effect from the extrapolation with a larger increase in uncertainty.

5. Discussion

We have presented a Bayesian nonparametric approach to the analysis of quantal data from bioassay studies. The modeling approach is built from nonparametric mixing driven by a simplified version of a dependent Dirichlet process prior, the linear-DDP prior. The implied dose-response curve is increasing, the inversion inferential objective becomes straightforward to address, and the calibration objective can be implemented without immense computational complexity. We have provided illustration with a commonly used data example which bears evidence of multiple modes and skewness in the dose-response curve. Further demonstration through portions of a cytogenetic

dosimetry data set reveals the practicality of the model.

Note that, under the single-p DDP mixture modeling approach discussed in Section 2.1, if the collection of mixing distributions $\{G_x : x \in \mathcal{X}\}$ is stochastically ordered in x, then $\Pr(y = 1 \mid G_x) = \int \Phi(\theta) \mathrm{d}G_x(\theta)$ is increasing in x. This provides a general sufficient condition to create mixture models that support monotonic dose-response curves through stochastic processes for the DDP atoms $\eta_{l\mathcal{X}}$ with increasing sample paths. Of methodological interest are extensions of the linear-DDP prior based on more general specifications for the DDP atoms, using structured Gaussian processes with increasing realizations. This extension is not likely to offer significant practical utility to the standard quantal bioassay setting where the entire dose-response curve is typically observed. However, it has potential for more flexible calibration inference in cytogenetic dosimetry where extrapolation for the dose-response relationship is necessary.

The cytogenetic dosimetry application provides the motivation for another extension of the methodology to modeling for dose-response studies with outcomes comprising an ordinal classification, as in the original experiment considered in Section 4.2. In this context, fully nonparametric mixture modeling is particularly attractive as it enables different shapes for the distinct dose-response curves, and a more practical modeling approach for a moderate to large number of ordinal categories than the model in Kottas et al. (2002) based on stochastically ordered DP priors.

6. Supplementary Materials

Web Appendix A referenced in Section 4, and the Fortran 90 code for implementing the MCMC algorithm presented in Section 3.1 are available with this paper at the *Biometrics* website on Wiley Online Library.

ACKNOWLEDGEMENTS

This research is part of the Ph.D. dissertation of Kassandra Fronczyk, completed at University of California, Santa Cruz, and was supported in part by the National Science Foundation under award DMS 1310438. The authors wish to thank the Associate Editor and two reviewers for their comments and feedback.

References

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

Ashford, J. R. and Walker, P. J. (1972). Quantal response analysis for mixture populations. *Biometrics* **28,** 981–988.

Barrientos, A. F., Jara, A., and Quintana, F. A. (2012). On the support of MacEachern's dependent Dirichlet processes and extensions. *Bayesian Analysis* **7,** 277–310.

Basu, S. and Mukhopadyay, S. (2000). Binary response regression with normal scale mixture links. In Dey, D., Ghosh, S., and Mallick, B., editors, *Generalized Linear Models: A Bayesian Perspective*, pages 231–241. Marcel Dekker, New York.

Bhattacharya, P. K. (1981). Posterior distribution of a Dirichlet process from quantal response data. *The Annals of Statistics* **9,** 803–811.

Bornkamp, B. and Ickstadt, K. (2009). Bayesian nonparametric estimation of continuous monotone functions with applications to dose-response analysis. *Biometrics* **65,** 198–205.

Casanova, M. P., Iglesias, P., and Bolfarine, H. (2010). A Bayesian semiparametric approach for solving the discrete calibration problem. *Communications in Statistics - Simulation and Computation* **39,** 347–360.

DeIorio, M., Johnson, W. O., Müller, P., and Rosner, G. L. (2009). Bayesian nonparametric non-proportional hazards survival modelling. *Biometrics* **63,** 762–771.

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

Disch, D. (1981). Bayesian nonparametric inference for effective doses in a quantal-response experiment. *Biometrics* **37,** 713–722.

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

Fronczyk, K. and Kottas, A. (2013). A Bayesian nonparametric modeling framework for developmental toxicity studies. *Journal of the American Statistical Association* To appear.

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

Geweke, J. and Keane, M. (1999). Mixture of normals probit models. In Hsiao, C., Pesaran, M. H., Lahiri, K., and Lee, L. F., editors, *Analysis of Panels and Limited Dependent Variable Models*, pages 49–78. Cambridge University Press.

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

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

Jara, A., García-Zattera, M. J., and Lesaffre, E. (2007). A Dirichlet process mixture model for the analysis of correlated binary responses. *Computational Statistics & Data Analysis* **51,** 5402–5415.

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

Kottas, A. and Fronczyk, K. (2013). Flexible Bayesian modelling for clustered categorical responses in developmental toxicology. In Damien, P., Dellaportas, P., Polson, N., and Stephens, D., editors, *Bayesian Theory and Applications*, pages 70–83. Oxford University Press.

Kuo, L. (1983). Bayesian bioassay design. *The Annals of Statistics* **11,** 886–895.

Kuo, L. (1988). Linear Bayes estimators of the potency curve in bioassay. *Biometrika* **75,** 91–96.

Lwin, T. and Martin, P. J. (1989). Probits of mixtures. *Biometrics* **45,** 721–732.

MacEachern, S. N. (1998). Computational methods for mixture of Dirichlet process models. In Dey, D., Müller, P., and Sinha, D., editors, *Practical Nonparametric and Semiparametric Bayesian Statistics*, pages 23–44. Springer-Verlag.

MacEachern, S. N. (2000). Dependent Dirichlet processes. Technical report, Ohio State University, Deptartment of Statistics.

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

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

Muliere, P. and Walker, S. (1997). A Bayesian nonparametric approach to determining a maximum tolerated dose. *Journal of Statistical Planning and Inference* **61,** 339–353.

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

Walker, P. J. (1966). A method of measuring the sensitivity of trypanosomes to acriflavine and trivalent tryparsamide. *Journal of General Microbiology* **43,** 45–58.

Table 1: Trypanosome data reported in Ashford and Walker (1972).

Log concentration	4.7	4.8	4.9	5.0	5.1	5.2	5.3	5.4
Exposed	55	49	60	55	53	53	51	50
Dead	0	8	18	18	22	37	47	50

Table 2: Cytogenetic dosimetry data. The cell counts for the binucleated cells from healthy, older subjects, as reported in Madruga et al. (1996).

^{60}Co radiation	$\geq$ 2 MN	1 MN	0 MN
0 cGy	2	31	920
20 cGy	8	41	989
50 cGy	14	56	933
100 cGy	32	114	939
200 cGy	67	176	794
300 cGy	59	209	683
400 cGy	107	256	742
500 cGy	143	327	771

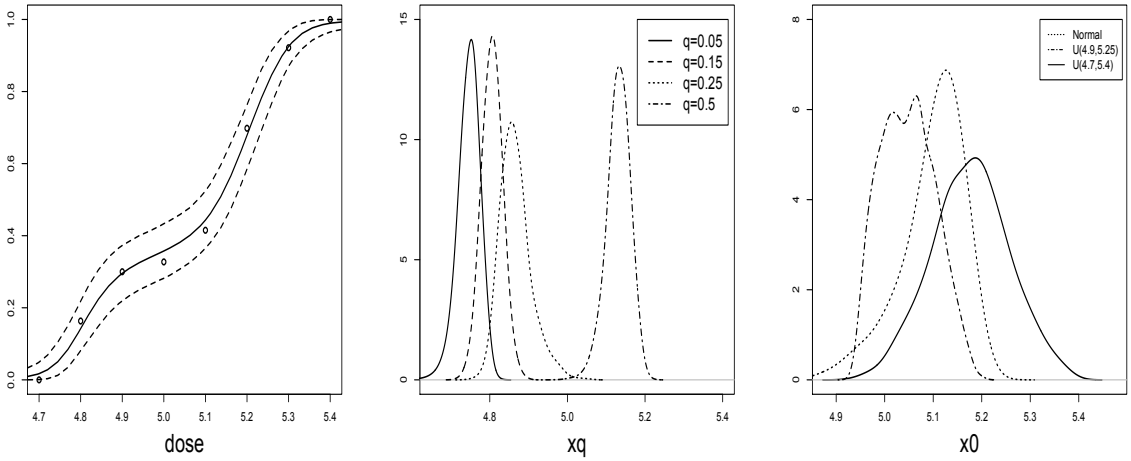

Figure 1: Trypanosome data. Left panel: posterior mean estimate (solid line) and 90% uncertainty bands (dashed lines) for the dose-response curve. Middle panel: inversion estimates x_q for $q = 0.05, 0.15, 0.25, 0.5$. Right panel: posterior densities for the calibrated dose level, given $n_0 = 52$ and $\sum_{j=1}^{n_0} y_{0j} = 26$, under a $\mathrm{N}(5.1, 0.07^2)$ prior (dotted line), a uniform prior on $(4.9, 5.25)$ (dashed-dotted line), and a uniform prior across the full dose range (solid line).

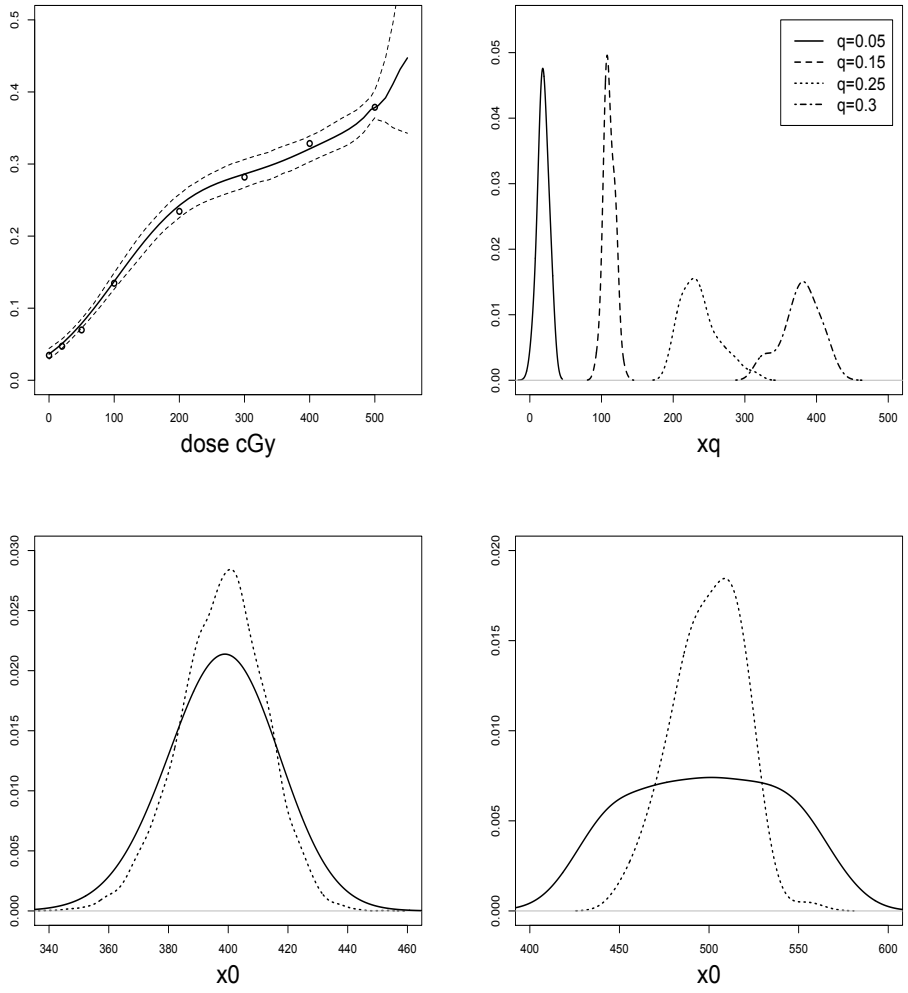

Figure 2: Cytogenetic dosimetry data. Top left panel: posterior mean estimate (solid line) and 90% uncertainty bands (dashed lines) for the dose-response curve. Top right panel: inversion estimates x_q for $q = 0.05, 0.15, 0.25, 0.3$. Bottom left panel: posterior densities for the calibrated dose level given the observed vector at dose 400 cGy including the data vector (dotted line) and leaving the vector out (solid line). Bottom right panel: posterior densities for the calibrated dose level given the observed vector at dose 500 cGy including the data vector (dotted line) and leaving the vector out (solid line).