AMS 241: Bayesian Nonparametric Methods
Notes 3 – Dependent nonparametric prior models

Instructor: Athanasios Kottas

Department of Applied Mathematics and Statistics
University of California, Santa Cruz
Fall 2015

Outline

1. Dependent Dirichlet processes
2. Nonparametric prior models for finite collections of distributions
3. Spatial Dirichlet process models
4. Application: DDP modeling for developmental toxicity studies

Dependent Dirichlet processes

- So far we have focused on problems where a single (possibly multivariate) distribution is assigned a nonparametric prior. This is consistent with the earlier developments in the Bayes nonparametrics literature.

- However, in many applications, the objective is modeling a collection of distributions $\mathcal{G} = \{G_{\mathbf{s}} : \mathbf{s} \in S\}$, indexed by $\mathbf{s} \in S$ — for example, S might be a discrete, finite set indicating different "groups", a time interval, a spatial region, or a covariate space.

- Obvious options:
  - Assume that the distribution is the same everywhere, e.g., $G_{\mathbf{s}} \equiv G \sim \text{DP}(\alpha, G_0)$ for all $\mathbf{s}$. This is too restrictive!
  - Assume that the distributions are independent and identically distributed, e.g., $G_{\mathbf{s}} \sim \text{DP}(\alpha, G_0)$ independently for each $\mathbf{s}$. This is wasteful!
- We would like something in between.

Dependent Dirichlet processes

- A similar dilemma arises in parametric models. Recall the random intercepts model:

$$y_{ij} = \theta_i + \epsilon_{ij}, \qquad \epsilon_{ij} \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2),$$
$$\theta_i = \eta + \nu_i, \qquad \nu_i \overset{i.i.d.}{\sim} \mathrm{N}(0, \tau^2),$$

  with $\eta \sim \mathrm{N}(\eta_0, \kappa^2)$.
  - If $\tau^2 \to 0$, we have $\theta_i = \eta$ for all i, i.e., all means are the same. "Maximum" borrowing of information across groups.
  - If $\tau^2 \to \infty$, all the means are different (and independent from each other). No information is borrowed.
- In a traditional random effects model, estimating τ^2 provides something in between (some borrowing of information across effects).
- How can we generalize this idea to distributions?
  - Note that a nonparametric specification for the random effects distribution is not enough, as the distribution of the errors is still Gaussian.

Modeling dependence in collections of random distributions

- A number of modeling approaches have been presented in the literature, including:
  - Introducing dependence through the baseline distributions of conditionally independent nonparametric priors: for example, product of mixtures of DPs (refer to Notes 1). Simple but restrictive.
  - Structured priors for a finite number of distributions through linear combinations of realizations from independent DPs (e.g., Müller et al., 2004; Kolossiatis et al., 2013).
  - Hierarchical nonparametric prior models for finite collections of distributions (Analysis of densities model, ANOVA DDP, hierarchical DP, nested DP) — discussed later in this set of notes.
  - Dependent Dirichlet process (DDP): Starting with the stick-breaking construction of the DP, and replacing the weights and/or atoms with appropriate stochastic processes on S (MacEachern, 1999; 2000). Very general procedure, most of the models discussed here can be framed as DDPs.

Definition of the dependent Dirichlet process

- Recall the constructive definition of the Dirichlet process: $G \sim \text{DP}(\alpha, G_0)$ if and only if

$$G = \sum_{\ell=1}^{\infty} \omega_\ell \delta_{\theta_\ell},$$

where the θ_ℓ are i.i.d. from G_0, and $\omega_1 = z_1$, $\omega_\ell = z_\ell \prod_{r=1}^{\ell-1}(1 - z_r)$, $\ell = 2, 3, \ldots$, with z_r i.i.d. Beta$(1, \alpha)$.

- To construct a DDP prior for the collection of random distributions, $\mathcal{G} = \{G_{\mathbf{s}} : \mathbf{s} \in S\}$, define $G_{\mathbf{s}}$ as

$$G_{\mathbf{s}} = \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{s}) \delta_{\theta_\ell(\mathbf{s})},$$

  - with $\{\theta_\ell(\mathbf{s}) : \mathbf{s} \in S\}$, for $\ell = 1, 2, \ldots$, independent realizations from a (centering) stochastic process $G_{0,S}$ defined on S
  - and stick-breaking weights defined through independent realizations $\{z_r(\mathbf{s}) : \mathbf{s} \in S\}$, $r = 1, 2, \ldots$, from a stochastic process on S with marginals $z_r(\mathbf{s}) \sim \text{Beta}(1, \alpha(\mathbf{s}))$ (or with common $\alpha(\mathbf{s}) \equiv \alpha$).

Dependent Dirichlet processes

- For any fixed $\mathbf{s}$, this construction yields a DP prior for distribution $G_{\mathbf{s}}$.
- The support of DDP priors is studied in Barrientos et al. (2012).
- For uncountable index sets S, smoothness (e.g., continuity) properties of the centering process $G_{0,S}$ and the stochastic process that defines the weights drive *smoothness* of DDP realizations.
  - For instance, for spatial regions S, we typically seek smooth evolution for the distributions $G_{\mathbf{s}}$, with the level of dependence between $G_{\mathbf{s}}$ and $G_{\mathbf{s}'}$ driven by the distance between spatial sites $\mathbf{s}$ and $\mathbf{s}'$.
- For specified set A, $\{G_{\mathbf{s}}(A) : \mathbf{s} \in S\}$ is a stochastic process with beta marginals. The covariance between $G_{\mathbf{s}}(A)$ and $G_{\mathbf{s}'}(A)$ can be used to study the dependence structure under a particular DDP prior.
- Effective inference under DDP prior models requires some form of replicate responses across the observed index points.
- As with DP priors, we usually employ the DDP prior to model the distribution of the parameters in a hierarchical model, resulting in DDP mixture models.

"Common-weights" dependent Dirichlet processes

- "Common-weights" (or "single-p") DDP models: the weights do not depend on $\mathbf{s}$; dependence is induced only from dependence across atoms in the stick-breaking construction:

$$G_{\mathbf{s}} = \sum_{\ell=1}^{\infty} \omega_\ell \delta_{\theta_\ell(\mathbf{s})}$$

where $\omega_1 = z_1$, $\omega_\ell = z_\ell \prod_{r=1}^{\ell-1}(1-z_r)$, $\ell \geq 2$, with z_r i.i.d. Beta$(1, \alpha)$.

  - Advantage $\Rightarrow$ Computation is relatively simple, since common-weights DDP mixture models can be written as DP mixtures for an appropriate baseline distribution.
  - Disadvantage $\Rightarrow$ Dependent weights can generate local dependence structure which is desirable in temporal or spatial applications.

- Some applications of common-weights DDP models: De Iorio et al. (2004); Rodriguez and ter Horst (2008); De Iorio et al. (2009); Di Lucca et al. (2013); Fronczyk and Kottas (2014a,b).

"Common-atoms" dependent Dirichlet processes

- "Common-atoms" DDP models: the alternative simplification where the atoms are common to all distributions:

$$G_{\mathbf{s}} = \sum_{\ell=1}^{\infty} \omega_\ell(\mathbf{s})\delta_{\theta_\ell}$$

where the θ_ℓ are i.i.d. from G_0.

  - Advantage $\Rightarrow$ The structure with common atoms across distributions that have weights that change with $\mathbf{s}$ may be attractive in certain applications. When the dimension of θ is moderate to large, it also reduces significantly the number of stochastic processes over S required for a full DDP specification.
  - Disadvantage $\Rightarrow$ Prediction at new $\mathbf{s}$ (say, forecasting when $\mathbf{s}$ corresponds to discrete time) can be problematic.

- Examples of modeling with common-atoms DDP priors: Taddy (2010) and Nieto-Barajas et al. (2012).

Dependent Dirichlet processes

- Section 2 provides an overview of three classes of nonparametric priors for finite collections of distributions: ANOVA DDP (De Iorio et al., 2004); hierarchical DPs (Teh. et al., 2006), which are related to the "analysis of densities" model (Tomlinson and Escobar, 1999); and nested DPs (Rodriguez et al., 2008).
- Section 3 presents spatial DPs (Gelfand et al., 2005; Kottas et al., 2008), and Section 4 an application of DDP modeling for risk assessment in developmental toxicity studies (Fronczyk and Kottas, 2014a).

- However, this is by no means an exhaustive list: order-depedent DDPs (Griffin and Steel, 2006); generalized spatial DP (Duan, Guindani and Gelfand, 2007); kernel stick-breaking processes (Dunson and Park, 2008); dependent Pólya tree regression models (Trippa et al., 2011); stick-breaking autoregressive processes (Griffin and Steel, 2011); .....

ANOVA dependent Dirichlet process models

- Consider a space S such that $\mathbf{s} = (s_1, \ldots, s_p)$ corresponds to a vector of categorical variables. For instance, in a clinical setting, G_{s_1,s_2} might correspond to the random effects distribution for patients treated at levels s_1 and s_2 of two different drugs.
- For example, define $y_{s_1,s_2,k} \mid G_{s_1,s_2}, \sigma^2 \sim \int \mathrm{N}(y_{s_1,s_2,k} \mid \eta, \sigma^2) \mathrm{d}G_{s_1,s_2}(\eta)$ where

$$G_{s_1,s_2} = \sum_{h=1}^{\infty} \omega_h \delta_{\theta_{h,s_1,s_2}}$$

with $\theta_{h,s_1,s_2} = m_h + A_{h,s_1} + B_{h,s_2} + AB_{h,s_1,s_2}$ and

$$m_h \sim G_0^m, \quad A_{h,s_1} \sim G_0^A, \quad B_{h,s_2} \sim G_0^B, \quad AB_{h,s_1,s_2} \sim G_0^{AB}.$$

- Typically G_0^m, G_0^A, G_0^B and G_0^{AB} are normal distributions and we introduce identifiability constrains such as $A_{h,1} = B_{h,1} = 0$ and $AB_{h,1,s_2} = AB_{h,s_1,1} = 0$.

ANOVA dependent Dirichlet process models

- Note that the atoms of G_{s_1,s_2} have a structure that resembles a two way ANOVA.
- Indeed, the ANOVA-DDP mixture model can be reformulated as a DP mixture of ANOVA models where, at least in principle, there can be up to one different ANOVA for each observation:

$$y_{s_1,s_2,k} \mid F, \sigma^2 \sim \int \mathrm{N}(y_{s_1,s_2,k} \mid d_{s_1,s_2}\eta, \sigma^2)\mathrm{d}F(\eta), \quad F \sim \mathrm{DP}(\alpha, G_0),$$

  where d_{s_1,s_2} is a design vector selecting the appropriate coefficients from η and $G_0 = G_0^m G_0^A G_0^B G_0^{AB}$.
- In practice, just a small number of ANOVA models. If a single component is used, we recover a parametric ANOVA model.
- Rephrasing the ANOVA-DDP model as a DP mixture simplifies posterior simulation.
  - Function `LDDPdensity` in `DPpackage` can be used to fit ANOVA-DDP models.

Hierarchical Dirichlet processes

- Consider modeling the distribution of SAT scores on different schools.
- Data y_{ij} might correspond to the SAT score obtained by student $j = 1, \ldots, m_i$ in school $i = 1, \ldots, n$.
- Traditionally, this type of data has been modeled using a random intercept model.

$$y_{ij} \mid \theta_i \sim \mathrm{N}(\theta_i, \sigma^2), \qquad \theta_i \mid \mu \sim \mathrm{N}(\mu, \tau^2), \qquad \mu \sim \mathrm{N}(\mu_0, \kappa^2),$$

  where θ_i is the school-specific random effect.

- But, what if the distribution of scores within a school appears to be (highly) non-Gaussian?

Hierarchical Dirichlet processes

- Hierarchical Dirichlet process (HDP) mixture models allow us to estimate the school-specific distribution by identifying latent classes of students that appear (possibly with different frequencies) in all schools.
- Let

$$y_{ij} \mid G_i \sim \int k(y_{ij} \mid \eta) \mathrm{d} G_i(\eta), \quad G_i \mid G_0 \sim \mathrm{DP}(\alpha, G_0), \quad G_0 \sim \mathrm{DP}(\beta, H).$$

- Conditionally on G_0, the mixing distribution for each school is an independent sample from a DP — dependence across schools is introduced, since they all share the same baseline measure G_0.
- This structure is reminiscent of the Gaussian random effects model, but it is built at the level of the distributions.

Hierarchical Dirichlet processes

- Since G_0 is drawn from a DP, it is (a.s.) discrete, $G_0 = \sum_{\ell=1}^{\infty} \omega_\ell \delta_{\phi_\ell}$.
- Therefore, when we draw the atoms for G_i we are forced to choose among $\phi_1, \phi_2, \ldots$, i.e., we can write G_i as:

$$G_i = \sum_{\ell=1}^{\infty} \pi_{\ell i} \delta_{\phi_\ell}$$

- Note that the weights assigned to the atoms *are not independent*. Intuitively, if ϕ_ℓ has a large weight ω_ℓ under G_0, then the weight $\pi_{\ell i}$ under G_i will likely be large for every i. Indeed, $\boldsymbol{\pi}_i = (\pi_{1i}, \pi_{2i}, \ldots) \sim \text{DP}(\alpha, \boldsymbol{\omega})$, where $\boldsymbol{\omega} = \{\omega_\ell : \ell = 1, 2, ...\}$, such that $\text{E}(\pi_{\ell i} \mid \boldsymbol{\omega}) = \omega_\ell$.
- Note that the HDP is essentially a common-atoms DDP prior model.
- An MCMC sampler can be devised for posterior simulation by composing two Pólya urns, one built from (α, G_0) and one from (β, H) — the resulting MCMC algorithm is similar to the marginal sampler for DP mixture models, but bookkeeping is harder.

Nested Dirichlet Processes

- Also a model for exchangeable distributions — rather than borrowing strength by sharing clusters among all distributions, the nested DP (NDP) borrows information by clustering similar distributions.
- An example: assessment for quality of care in hospitals nationwide.
  - y_{ij}: percentage of patients in hospital $j = 1, \ldots, m_i$ within state $i = 1, \ldots, n$ who received the appropriate antibiotic on admission.
  - We may want to cluster states with similar distributions of quality scores, and simultaneously cluster hospitals with similar outcomes.

- Let $y_{ij} \mid G_i \sim \int k(y_{ij} \mid \eta) \mathrm{d} G_i(\eta)$, where

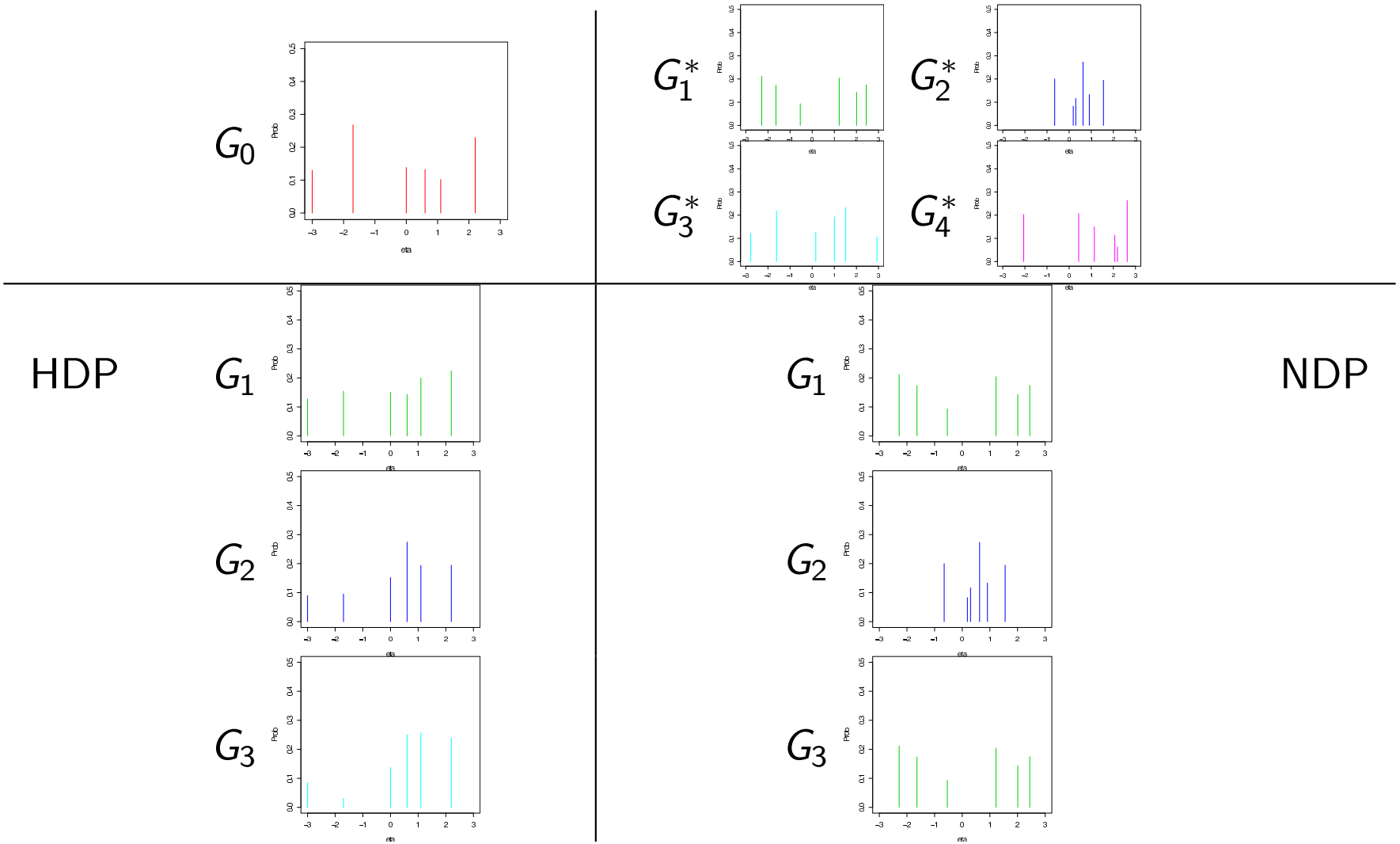
$$G_i \sim \sum_{k=1}^{K} \omega_k \delta_{G_k^*} \qquad\qquad G_k^* = \sum_{\ell=1}^{\infty} \pi_{\ell k} \delta_{\theta_{\ell k}},$$

where $\theta_{\ell k} \sim H$, $\pi_{\ell k} = u_{\ell k} \prod_{r<\ell}(1 - u_{rk})$ with $u_{\ell k} \sim \text{Beta}(1, \beta)$, and $\omega_k = v_k \prod_{r<k}(1 - v_r)$ with $v_k \sim \text{Beta}(1, \alpha)$.

Nested Dirichlet Processes

- In this case, we write $\{G_1, \ldots, G_n\} \sim \text{DP}(\alpha, \text{DP}(\beta, H))$.
- Note that the NDP generates two layers of clustering: states, and hospitals within groups of states. However, groups of states are conditionally independent from each other.
- The NDP is not a common-weights DDP model.
- A standard marginal sampler is not feasible in this problem — computation can be carried out using an extension of the blocked Gibbs sampler.

The HDP vs. the NDP



Spatial Dirichlet process models

- Spatial data modeling: based on **Gaussian processes** (distributional assumption) and **stationarity** (assumption on the dependence structure).
- Basic model for a spatial random field $\mathbf{Y}_D = \{Y(s) : s \in D\}$, with $D \subseteq R^d$:

$$Y(s) = \mu(s) + \theta(s) + \epsilon(s).$$

  - $\mu(s)$ mean process, e.g., $\mu(s) = x'(s)\beta$.
  - $\theta(s)$ a spatial process, typically, a mean 0 isotropic Gaussian process, i.e., $\text{Cov}(\theta(s_i), \theta(s_j) \mid \sigma^2, \phi) = \sigma^2 \rho_\phi(||s_i - s_j||) = \sigma^2 (H(\phi))_{i,j}$
  - A pure error (nugget) process, e.g., $\epsilon(s)$ i.i.d. $N(0, \tau^2)$.
- Induced model for observed sample (**point referenced spatial data**), $\mathbf{Y} = (Y(s_1), \ldots, Y(s_n))$, at sites $\mathbf{s}^{(n)} = (s_1, \ldots, s_n)$ in D

$$\mathbf{Y}|\beta, \sigma^2, \phi, \tau^2 \sim N(X'\beta, \sigma^2 H(\phi) + \tau^2 I_n).$$

Spatial Dirichlet process models

- **Objective of Bayesian nonparametric modeling**: develop prior models for the distribution of $\theta_D = \{\theta(s) : s \in D\}$, and thus for the distribution of $\mathbf{Y}_D = \{Y(s) : s \in D\}$, that relax the Gaussian **and** stationarity assumptions.

- In general, a fully nonparametric approach requires replicate observations at each site, $\mathbf{Y}_t = (Y_t(s_1), \ldots, Y_t(s_n))'$, $t = 1, \ldots, T$, though imbalance or missingness in the $Y_t(s_i)$ can be handled.

- Temporal replications available in various applications, e.g., in epidemiology, environmental contamination, and weather modeling.
  - Direct application of the methodology for spatial processes (when replications can be assumed approximately independent).
  - More generally, extension to **spatio-temporal modeling**, e.g., through dynamic spatial process modeling viewing $Y(s,t) \equiv Y_t(s)$ as a temporally evolving spatial process (Kottas, Duan and Gelfand, 2008).

Spatial Dirichlet process models

- **Spatial Dirichlet process**: arises as a dependent DP where G_0 is extended to G_{0D}, a random field over D, e.g., a stationary Gaussian process — thus, in the DP constructive definition, each θ_ℓ is extended to $\theta_{\ell,D} = \{\theta_\ell(s) : s \in D\}$ a realization from G_{0D}, i.e., a random surface over D.
- Hence, the spatial DP is defined as a random process over D

$$G_D = \sum_{\ell=1}^{\infty} \omega_\ell \delta_{\theta_{\ell,D}},$$

  which is centered at G_{0D}.
- A process defined in this way is denoted $G_D \sim \text{SDP}(\alpha, G_{0D})$.

Spatial Dirichlet process models

- Key property: if

$$\theta_D = \{\theta(s) : s \in D\} \mid G_D \sim G_D, \qquad G_D \sim \text{SDP}(\alpha, G_{0D})$$

then for any $\mathbf{s}^{(n)} = (s_1, \ldots, s_n)$, G_D induces $G^{(\mathbf{s}^{(n)})} \equiv G^{(n)}$, a random distribution for $(\theta(s_1), \ldots, \theta(s_n))$, and $G^{(n)} \sim \text{DP}(\alpha, G_0^{(n)})$, where $G_0^{(n)} \equiv G_0^{(\mathbf{s}^{(n)})}$.

- If G_{0D} is a Gaussian process, then $G_0^{(\mathbf{s}^{(n)})}$ is n-variate normal.

Spatial Dirichlet process models

- For stationary G_{0D}, the smoothness of realizations from SDP(α, G_{0D}) is determined by the choice of the covariance function of G_{0D}.
  - For instance, if G_{0D} produces a.s. continuous realizations, then $G^{(s)} - G^{(s')} \to 0$ a.s. as $||s - s'|| \to 0$.
  - We can learn about $G^{(s)}$ more from data at neighboring locations than from data at locations further away (as in usual spatial prediction).
- Random process G_D is centered at a stationary Gaussian process, but it is **nonstationary**, it has **nonconstant variance**, and it yields **non-Gaussian** finite dimensional distributions
- More general spatial DP models?
  - Allow weights to change with spatial location, i.e., allow realization at location s to come from a different surface than that for the realization at location s' (Duan, Guindani and Gelfand, 2007).

Spatial Dirichlet process models

- Almost sure discreteness of realizations from G_D?
  - Mix G_D against a pure error process $\mathcal{K}$ (i.i.d. $\epsilon(s)$ with mean 0 and variance τ^2) to create random process over D with continuous support.
- **Spatial DP mixture model:** If $G_D \sim \text{SDP}(\alpha, G_{0D})$, $\theta_D \mid G_D \sim G_D$, and $\mathbf{Y}_D - \theta_D \mid \tau^2 \sim \mathcal{K}$

$$F\left(\mathbf{Y}_D \mid G_D, \tau^2\right) = \int \mathcal{K}\left(\mathbf{Y}_D - \theta_D \mid \tau^2\right) \mathrm{d}G_D\left(\theta_D\right)$$

  i.e., $Y(s) = \theta(s) + \epsilon(s)$; $\theta(s)$ from a spatial DP; $\epsilon(s)$, say, i.i.d. $N(0, \tau^2)$ (again, process F is **non-Gaussian** and **nonstationary**).
- Adding covariates, the induced model at locations $\mathbf{s}^{(n)} = (s_1, \ldots, s_n)$,

$$f\left(\mathbf{Y} \mid G^{(n)}, \beta, \tau^2\right) = \int f_{N_n}\left(\mathbf{Y} \mid X'\beta + \theta, \tau^2 I_n\right) \mathrm{d}G^{(n)}\left(\theta\right),$$

  where $\mathbf{Y} = (Y(s_1), \ldots, Y(s_n))'$, $\theta = (\theta(s_1), \ldots, \theta(s_n))'$, and X is a $p \times n$ matrix with X_{ij} the value of the i-th covariate at the j-th location.

Spatial Dirichlet process models

- Data: for $t = 1, \ldots, T$, response $\mathbf{Y}_t = (Y_t(s_1), \ldots, Y_t(s_n))'$ (with latent vector $\theta_t = (\theta_t(s_1), \ldots, \theta_t(s_n))'$), and design matrix X_t.
- $G_0^{(n)}(\cdot \mid \sigma^2, \phi) = \mathrm{N}_n(\cdot \mid 0_n, \sigma^2 H_n(\phi))$ where $(H_n(\phi))_{i,j} = \rho_\phi(s_i - s_j)$ (or $\rho_\phi(||s_i - s_j||)$), induced by a mean 0 stationary (or isotropic) Gaussian process. (Exponential covariance function $\rho_\phi(|| \cdot ||) = \exp(-\phi|| \cdot ||)$, $\phi > 0$, used for the data example.)
- Bayesian model: (*conjugate* DP mixture model)

$$\begin{aligned}
\mathbf{Y}_t \mid \theta_t, \beta, \tau^2 &\overset{ind}{\sim} \mathrm{N}_n(\mathbf{Y}_t \mid X_t'\beta + \theta_t, \tau^2 I_n), && t = 1, \ldots, T, \\
\theta_t \mid G^{(n)} &\overset{iid}{\sim} G^{(n)}, && t = 1, \ldots, T, \\
G^{(n)} \mid \alpha, \sigma^2, \phi &\sim \mathrm{DP}(\alpha, G_0^{(n)}); G_0^{(n)} = N_n(\cdot \mid 0_n, \sigma^2 H_n(\phi)),
\end{aligned}$$

  with hyperpriors for β, τ^2, α,σ^2, and ϕ.
- Posterior inference using standard MCMC techniques for DP mixtures — extensions to accommodate missing data — methods for prediction at new spatial locations.

Data example

- Precipitation data from the Languedoc-Rousillon region in southern France.
- Data were discussed, for example, in Damian, Sampson and Guttorp (2001).
  - Original version of the dataset includes 108 altitude-adjusted 10-day aggregated precipitation records for the 39 sites in Figure 4.6.
- We work with a subset of the data based on the 39 sites but only 75 replicates (to avoid records with too many 0-s), which have been log-transformed with site specific means removed
- Preliminary exploration of the data suggests that spatial association is higher in the northeast than in the southwest.
- In the interest of validation for spatial prediction, we removed two sites from each of the three subregions in Figure 4.6, specifically, sites s_4, s_{35}, s_{29}, s_{30}, s_{13}, s_{37}, and refitted the model using only the data from the remaining 33 sites.

Data example

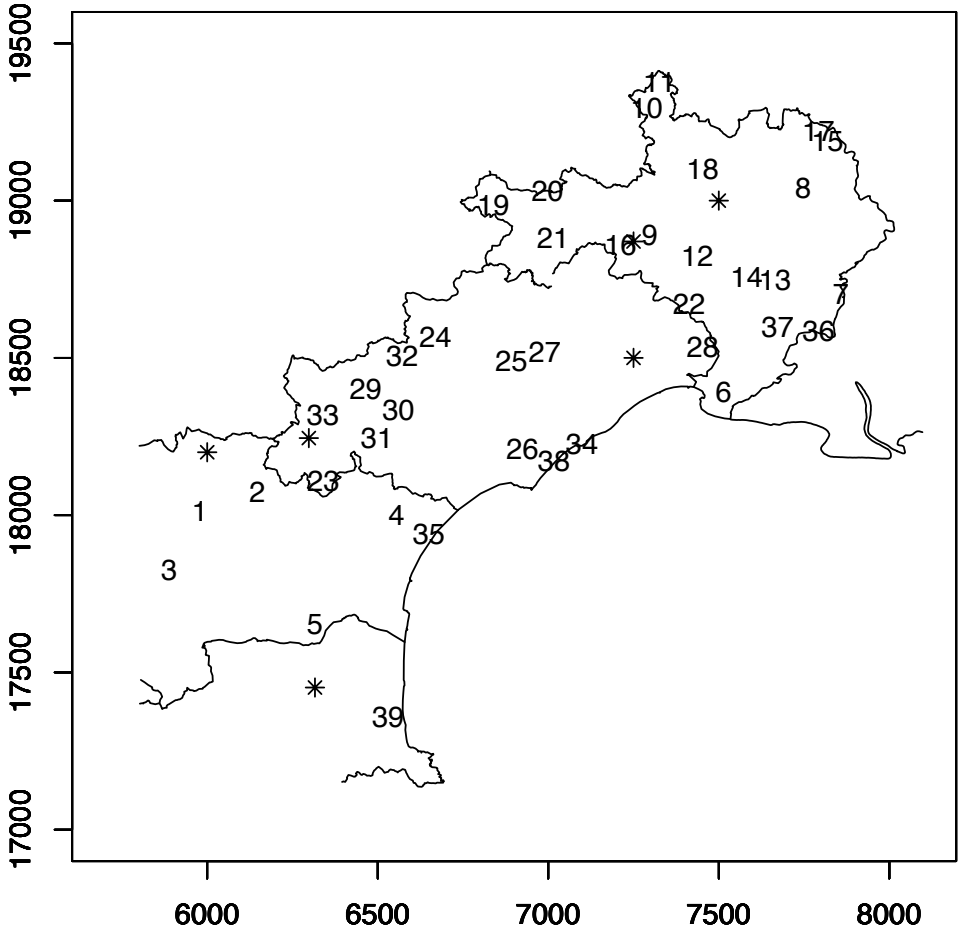


Figure 3.1: Geographic map of the Languedoc-Roussillon region in southern France.

Data example

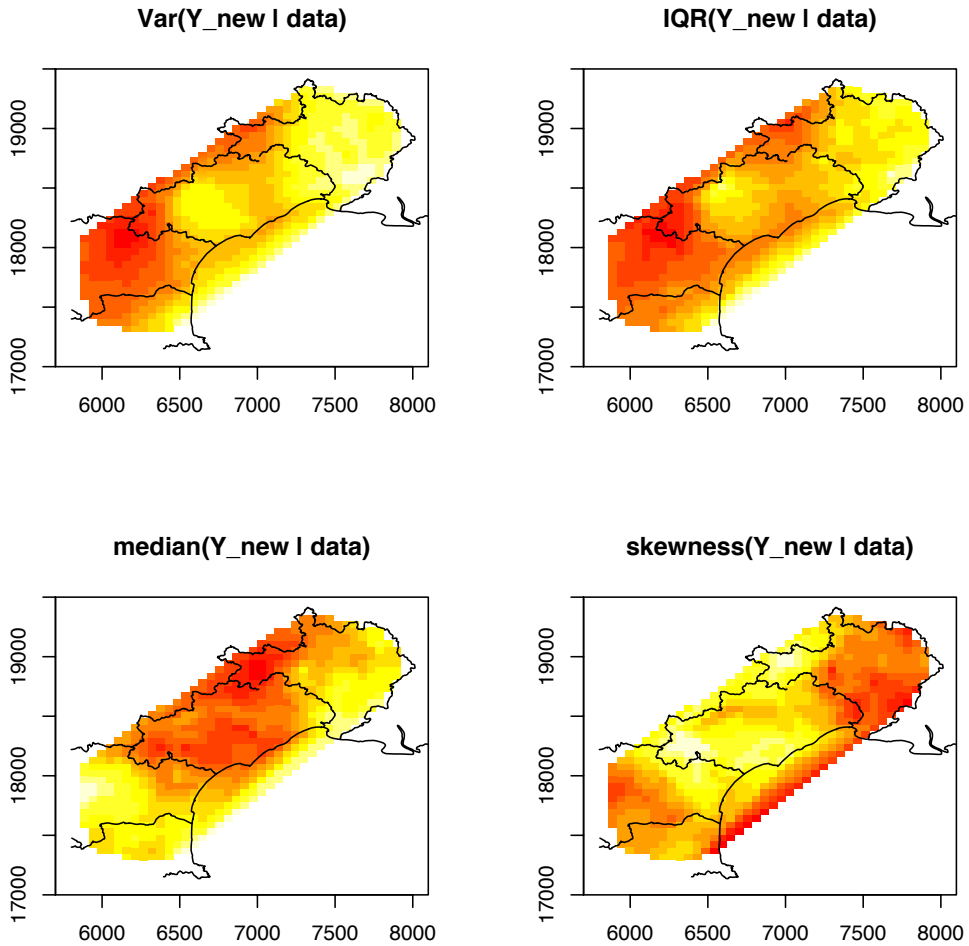


Figure 3.2: French precipitation data. Image plots based on functionals of posterior predictive distributions at observed sites and a number of new sites (darker colors correspond to smaller values).

Data example

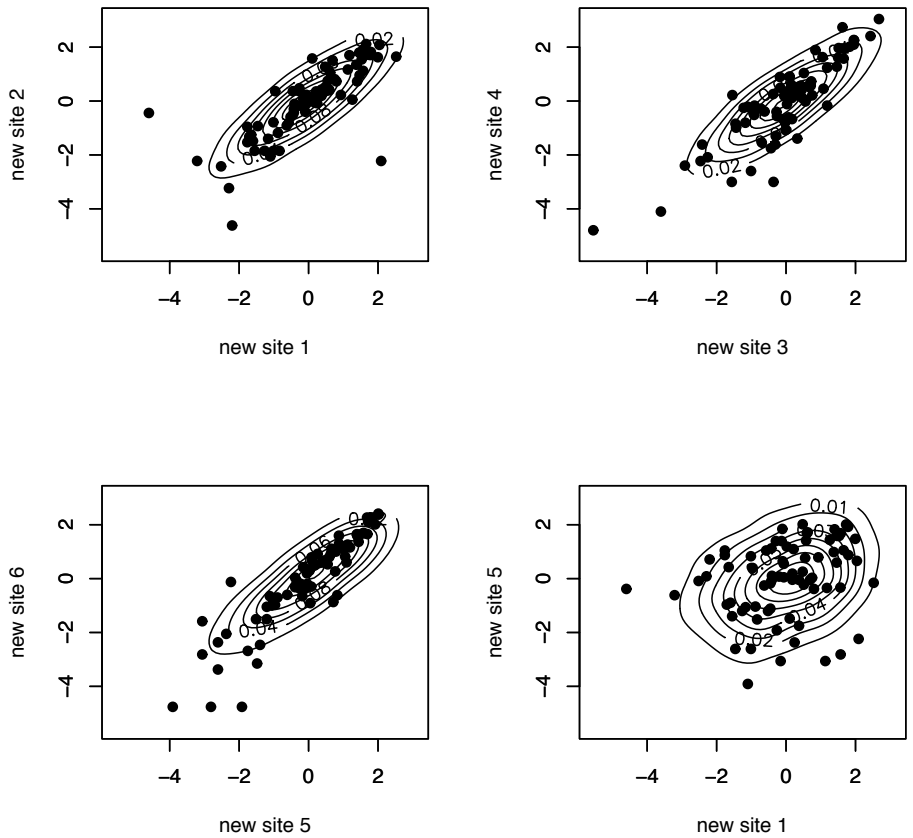


Figure 3.3: French precipitation data. Bivariate posterior predictive densities for pairs of sites (s_4, s_{35}), (s_{29}, s_{30}), (s_{13}, s_{37}) and (s_4, s_{13}) based on model fitted to data after removing sites s_4, s_{35}, s_{29}, s_{30}, s_{13} and s_{37} (overlaid on data observed at the corresponding pairs of sites in the full dataset).

DDP modeling for developmental toxicity studies

- Birth defects induced by toxic chemicals are investigated through developmental toxicity studies.
- A number of pregnant laboratory animals (dams) are exposed to a toxin. Recorded from each animal are:
  - the number of resorptions and/or prenatal deaths;
  - the number of live pups, and the number of live malformed pups;
  - data may also include continuous outcomes from the live pups (typically, body weight).
- Key objective is to examine the relationship between the level of exposure to the toxin (dose level) and the probability of response for the different endpoints: embryolethality; malformation; low birth weight.

Developmental toxicology data

- Focus on clustered categorical responses.
- Data structure for Segment II designs (exposure after implantation).
  - Data at dose (toxin) levels, x_i, $i = 1, ..., N$, including a control group (dose $= 0$).
  - n_i dams at dose level x_i.
  - For the j-th dam at dose x_i:
    - m_{ij}: number of implants.
    - R_{ij}: number of resorptions and prenatal deaths ($R_{ij} \leq m_{ij}$).
    - $\mathbf{y}_{ij}^* = \{y_{ijk}^* : k = 1, ..., m_{ij} - R_{ij}\}$: binary malformation indicators for the live pups ($y_{ij} = \sum_{k=1}^{m_{ij}-R_{ij}} y_{ijk}^*$: number of live pups with a malformation).

Developmental toxicology data

To begin with, consider simplest data form, $\{(m_{ij}, z_{ij}) : \ i = 1, \ldots, N, \ j = 1, \ldots, n_i\}$, where $z_{ij} = R_{ij} + y_{ij}$ is the number of combined negative outcomes

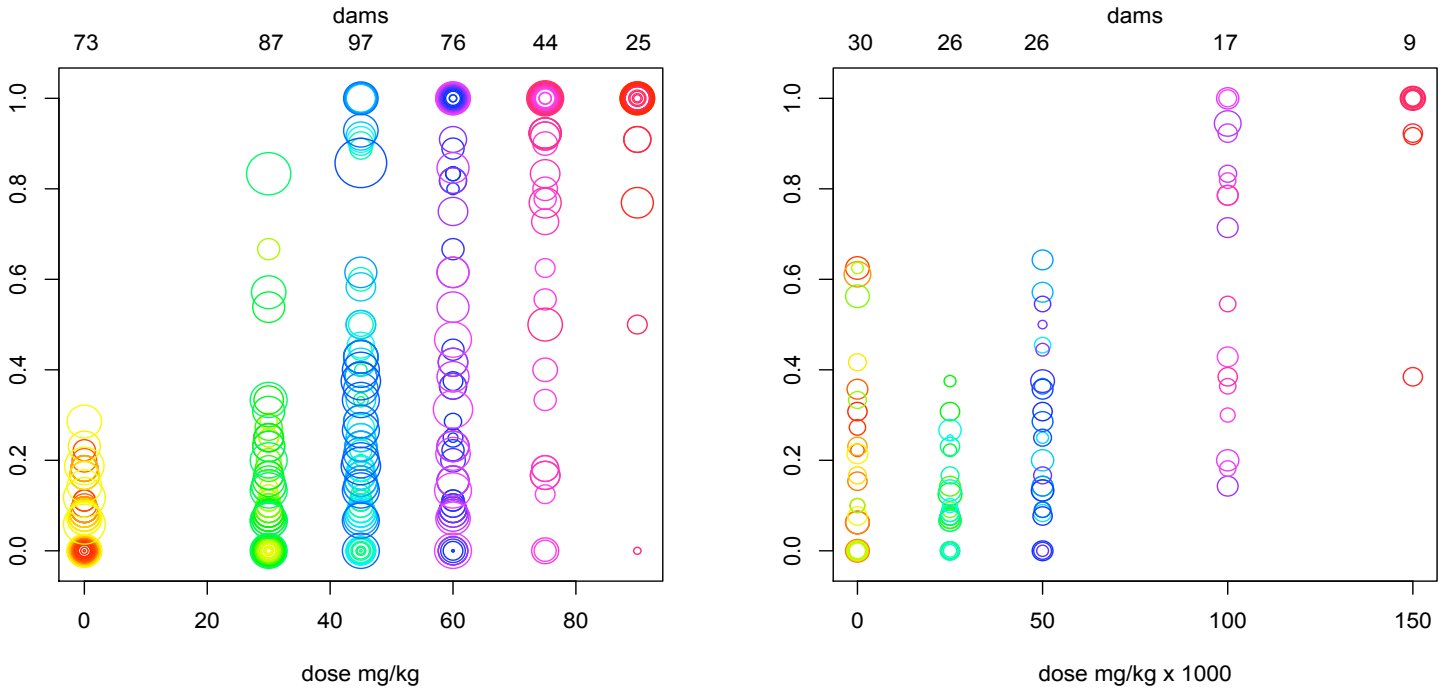


Figure 3.4: 2,4,5-T data (left) and DEHP data (right). Each circle is for a particular dam, the size of the circle is proportional to the number of implants, and the coordinates of the circle are the toxin level and the proportion of combined negative outcomes.

Objectives of DDP modeling

- Develop nonparametric Bayesian methodology for risk assessment in developmental toxicology.
  - Overcome limitations of parametric approaches, while retaining a fully inferential probabilistic model setting.
  - Modeling framework that provides flexibility in both the response distribution **and** the dose-response relationship.
- Build flexible risk assessment inference tools from nonparametric modeling for dose-dependent response distributions.
  - Nonparametric mixture models with increasing levels of complexity in the kernel structure to account for the different data types.
  - DDP priors for the dose-dependent mixing distributions.
  - Emphasis on properties of the implied dose-response relationships.

DDP mixture model formulation

- Begin with a DDP mixture model for the simplest data structure, $\{(m_{ij}, z_{ij}) : \ i = 1, \ldots, N, \ j = 1, \ldots, n_i\}$, where z_{ij} is the number of combined negative outcomes on resorptions/prenatal deaths and malformations.

- Number of implants is a random variable, though with no information about the dose-response relationship (the toxin is administered after implantation) — $f(m) = \text{Poisson}(m; \lambda)$, $m \geq 1$ (more general models can be used).

- Focus on dose-dependent conditional response distributions $f(z \mid m)$:

  - for dose level x, model $f(z \mid m) \equiv f(z \mid m; G_x)$ through a nonparametric mixture of Binomial distributions;
  - Common-weights DDP prior for the collection of mixing distributions $\{G_x : x \in \mathcal{X} \subseteq \mathbb{R}^+\}$.

DDP mixture model formulation

- DDP mixture of Binomial distributions:

$$f(z \mid m; G_{\mathcal{X}}) = \int \text{Bin}\left(z; m, \frac{\exp(\theta)}{1+\exp(\theta)}\right) \mathrm{d}G_{\mathcal{X}}(\theta), \quad G_{\mathcal{X}} \sim \text{DDP}(\alpha, G_{0\mathcal{X}})$$

- Gaussian process (GP) for $G_{0\mathcal{X}}$ with:
  - linear mean function, $\text{E}(\theta_\ell(x) \mid \beta_0, \beta_1) = \beta_0 + \beta_1 x$;
  - constant variance, $\text{Var}(\theta_\ell(x) \mid \sigma^2) = \sigma^2$;
  - isotropic power exponential correlation function, $\text{Corr}(\theta_\ell(x), \theta_\ell(x') \mid \phi) = \exp(-\phi|x - x'|^d)$ (with fixed $d \in [1, 2]$).
  - Hyperpriors for α and $\boldsymbol{\psi} = (\beta_0, \beta_1, \sigma^2, \phi)$.

- MCMC posterior simulation using blocked Gibbs sampling.
- Posterior predictive inference over observed and new dose levels, using the posterior samples from the model and GP interpolation for the DDP locations.

DDP mixture model formulation

- Key aspects of the DDP mixture model:
  - Flexible inference at each observed dose level through a nonparametric Binomial mixture (overdispersion, skewness, multimodality).
  - Prediction at unobserved dose levels (within and outside the range of observed doses).
  - Level of dependence between G_x and $G_{x'}$, and thus between $f(z \mid m; G_x)$ and $f(z \mid m; G_{x'})$, is driven by the distance between x and x'.
  - In prediction for $f(z \mid m; G_x)$, we learn more from dose levels x' nearby x than from more distant dose levels.
  - Inference for the dose-response relationship is induced by flexible modeling for the underlying response distributions.
- Linear mean function for the DDP centering GP enables connections with parametric models, and is key for flexible inference about the dose-response relationship.

Dose-response curve

- Exploit connection of the DDP Binomial mixture for the negative outcomes within a dam and a DDP mixture model with a product of Bernoullis kernel for the set of binary responses for all implants corresponding to that dam.

- Using the equivalent mixture model formulation for the underlying binary outcomes, define the dose-response curve as the probability of a negative outcome for a generic implant expressed as a function of dose level:

$$D(x) = \int \frac{\exp(\theta)}{1+\exp(\theta)} \mathrm{d}G_x(\theta) = \sum_{\ell=1}^{\infty} \omega_\ell \frac{\exp(\theta_\ell(x))}{1+\exp(\theta_\ell(x))}, \quad x \in \mathcal{X}$$

- If $\beta_1 > 0$, the prior expectation $\mathrm{E}(D(x))$ is non-decreasing with x, but prior (and thus posterior) realizations for the dose-response curve are not structurally restricted to be non-decreasing (a model asset!).

Data examples: 2,4,5-T data

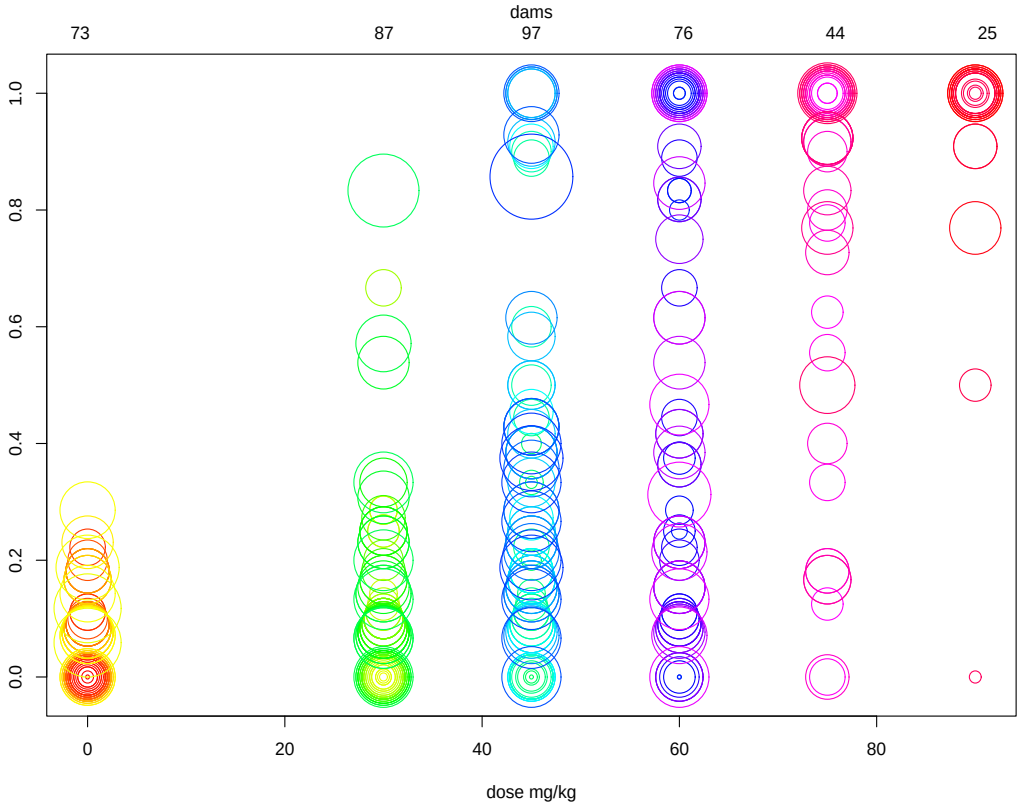


Figure 3.5: 2,4,5-T data. Data set from a developmental toxicity study regarding the effects of the herbicide 2,4,5-trichlorophenoxiacetic (2,4,5-T) acid.

Data examples: 2,4,5-T data

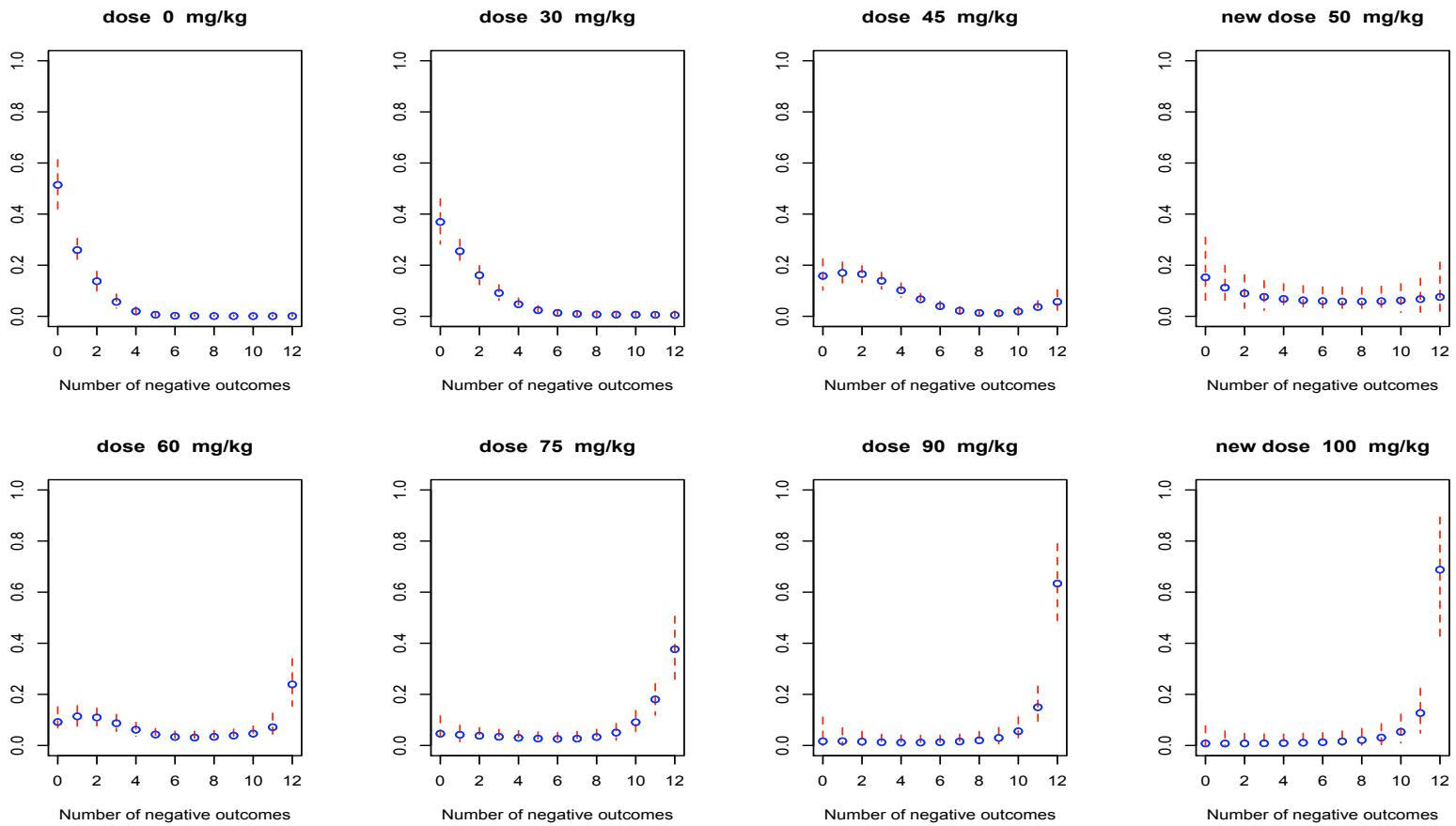


Figure 3.6: 2,4,5-T data. For the 6 observed and 2 new doses, posterior mean estimates (denoted by "o") and 90% uncertainty bands (red) for $f(z \mid m = 12; G_x)$.

Data examples: 2,4,5-T data

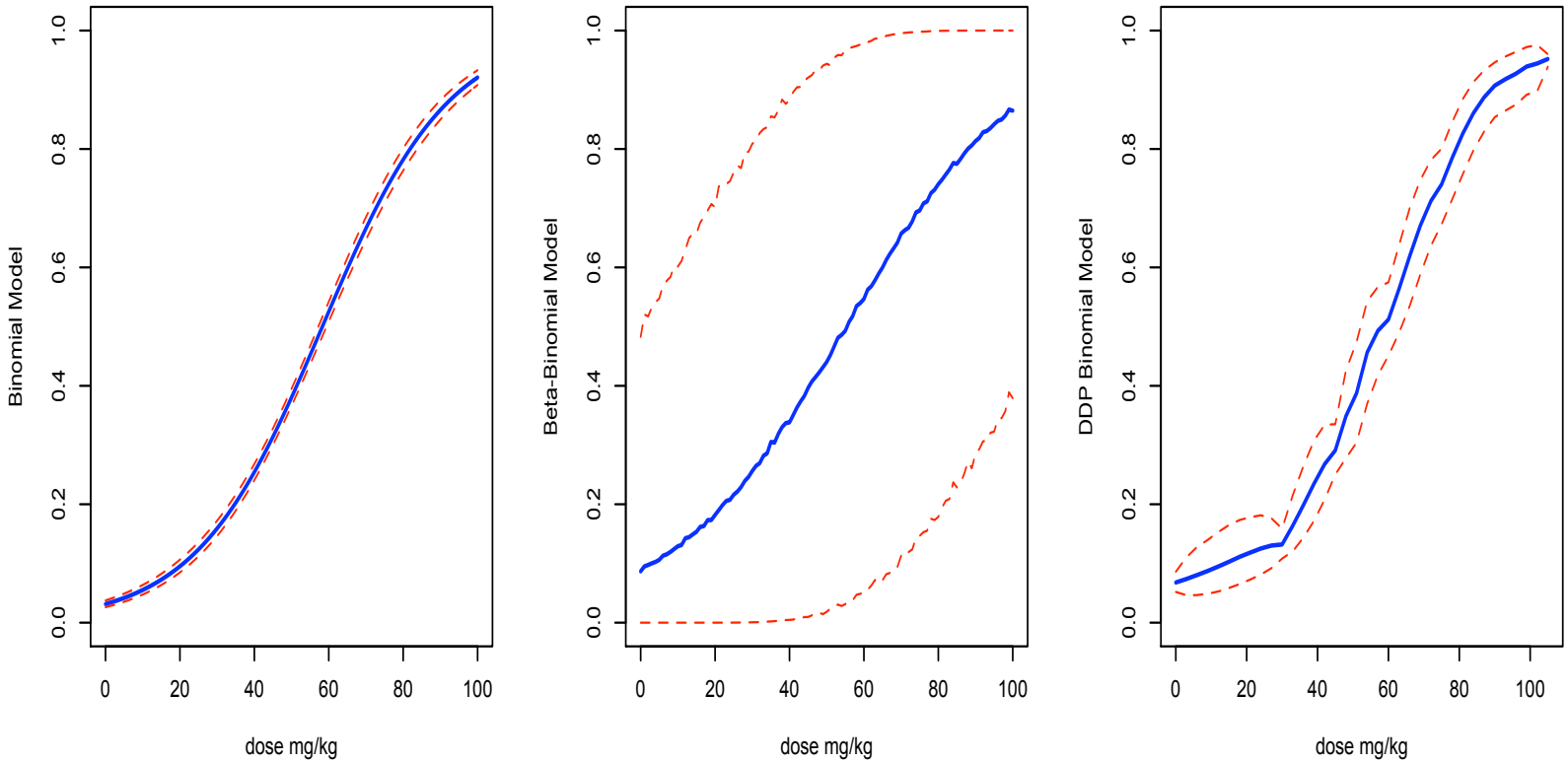


Figure 3.7: 2,4,5-T data. Posterior mean estimate and 90% uncertainty bands for the dose-response curve under a Binomial-logistic model (left), a Beta-Binomial model (middle), and the DDP Binomial mixture model (right).

Data examples: DEHP data

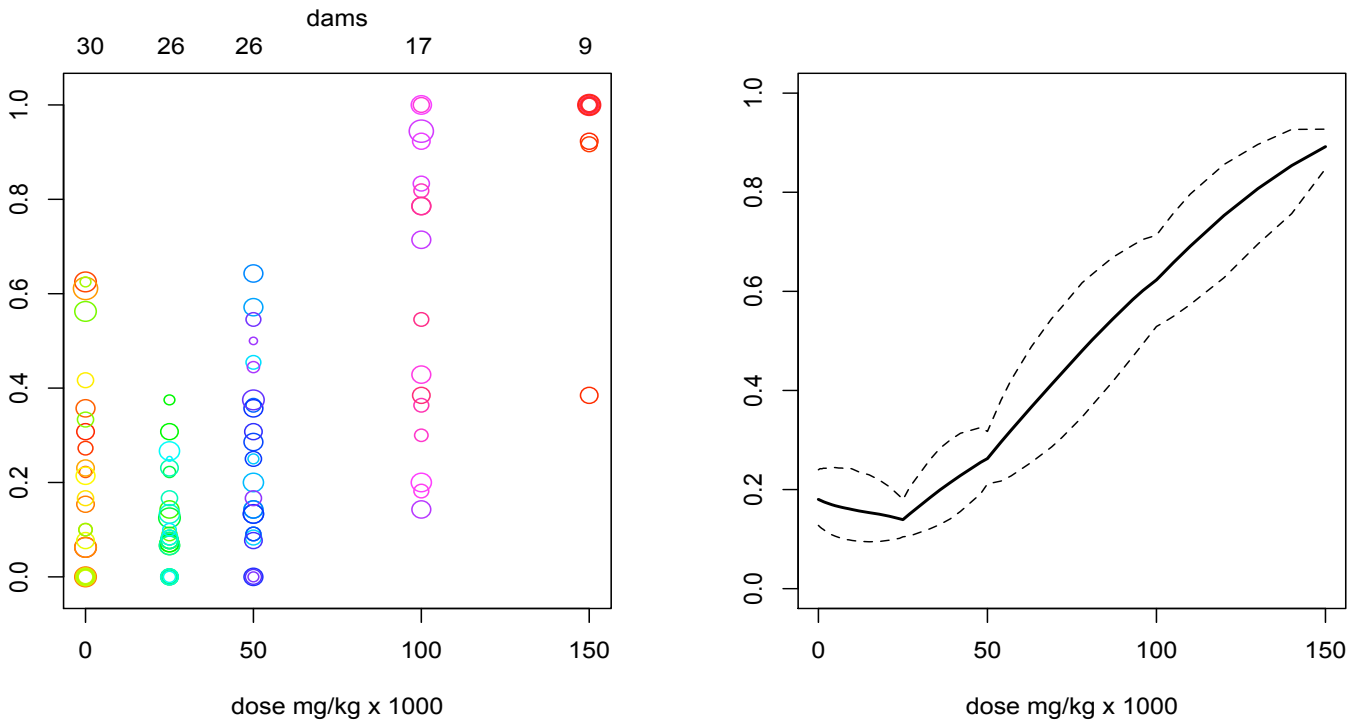


Figure 3.8: DEHP data. Left panel: data from an experiment that explored the effects of diethylhexalphthalate (DEHP), a commonly used plasticizing agent. Right panel: Posterior mean estimate and 90% uncertainty bands for the dose-response curve; the dip at small toxin levels may indicate a hormetic dose-response relationship.

Modeling for multicategory classification responses

Full version of the DEHP data

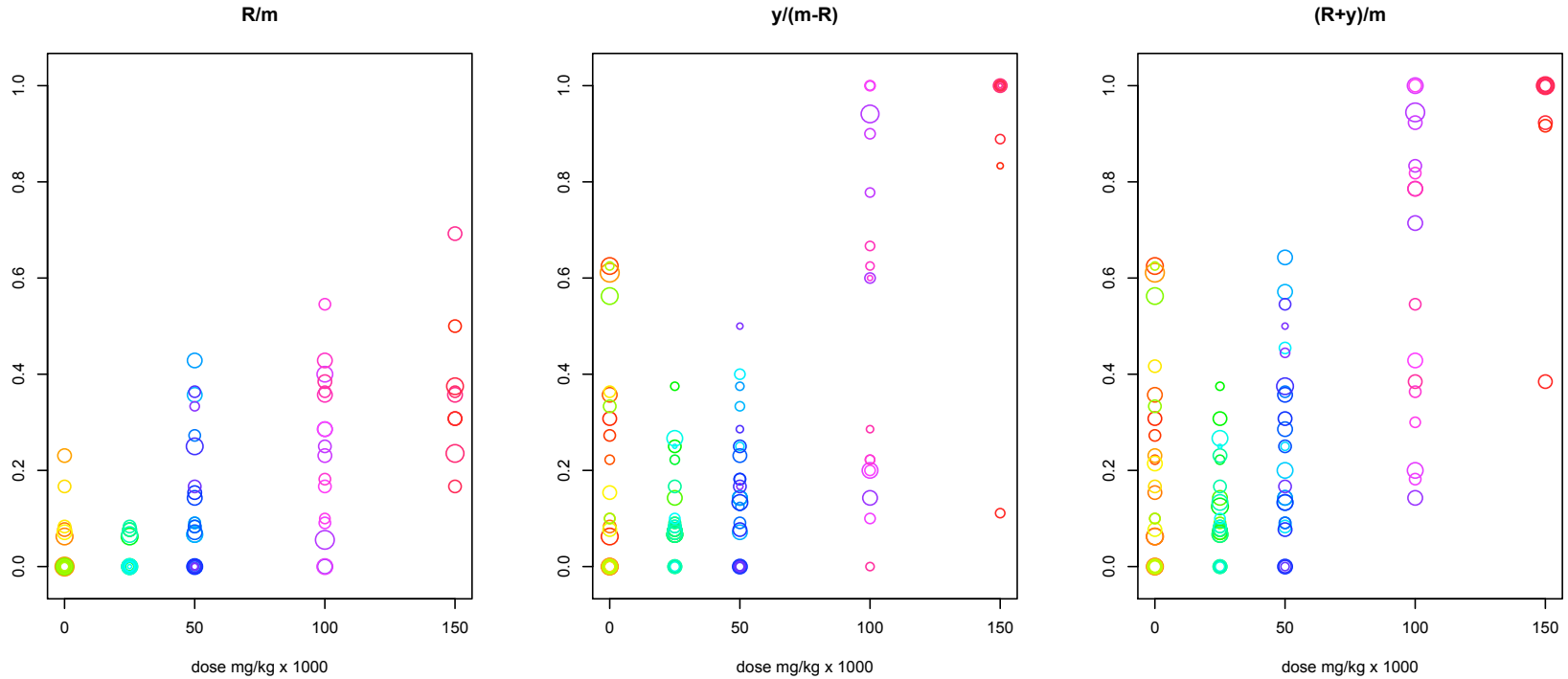


Figure 3.9: Clustered categorical responses: for the j-th dam at dose x_i, R_{ij} resorptions and prenatal deaths ($R_{ij} \leq m_{ij}$), and y_{ij} malformations among the live pups ($y_{ij} \leq m_{ij} - R_{ij}$).

Modeling for multicategory classification responses

- DDP mixture model for endpoints of embryolethality (R) and malformation for live pups (y)

$$f(R, y \mid m; G_{\mathcal{X}}) = \int \text{Bin}\left(R; m, \pi(\gamma)\right) \text{Bin}\left(y; m - R, \pi(\theta)\right) \mathrm{d}G_{\mathcal{X}}(\gamma, \theta)$$

  - $\pi(v) = \exp(v)/\{1 + \exp(v)\}$, $v \in \mathbb{R}$, denotes the logistic function;
  - $G_{\mathcal{X}} = \sum_{\ell=1}^{\infty} \omega_\ell \delta_{\eta_{\ell\mathcal{X}}} \sim \text{DDP}(\alpha, G_{0\mathcal{X}})$, where $\eta_\ell(x) = (\gamma_\ell(x), \theta_\ell(x))$;
  - $G_{0\mathcal{X}}$ defined through two independent GPs with linear mean functions, $\text{E}(\gamma_\ell(x) \mid \xi_0, \xi_1) = \xi_0 + \xi_1 x$, and $\text{E}(\theta_\ell(x) \mid \beta_0, \beta_1) = \beta_0 + \beta_1 x$.

- Equivalent mixture model (with product Bernoulli kernels) for binary responses: R^* non-viable fetus indicator; y^* malformation indicator.

Dose-response curves

- Probability of embryolethality:

$$\Pr(R^* = 1; G_x) = \int \pi(\gamma)\, \mathrm{d}G_x(\gamma, \theta), \quad x \in \mathcal{X}$$

(monotonic in prior expectation provided $\xi_1 > 0$).

- Probability of malformation:

$$\Pr(y^* = 1 \mid R^* = 0; G_x) = \frac{\int \{1 - \pi(\gamma)\}\pi(\theta)\, \mathrm{d}G_x(\gamma, \theta)}{\int \{1 - \pi(\gamma)\}\, \mathrm{d}G_x(\gamma, \theta)}, \quad x \in \mathcal{X}$$

- Combined risk function:

$$\Pr(R^* = 1 \text{ or } y^* = 1; G_x) = 1 - \int \{1-\pi(\gamma)\}\{1-\pi(\theta)\}\, \mathrm{d}G_x(\gamma, \theta), \quad x \in \mathcal{X}$$

(monotonic in prior expectation provided $\xi_1 > 0$ and $\beta_1 > 0$)

DEHP data (full version)

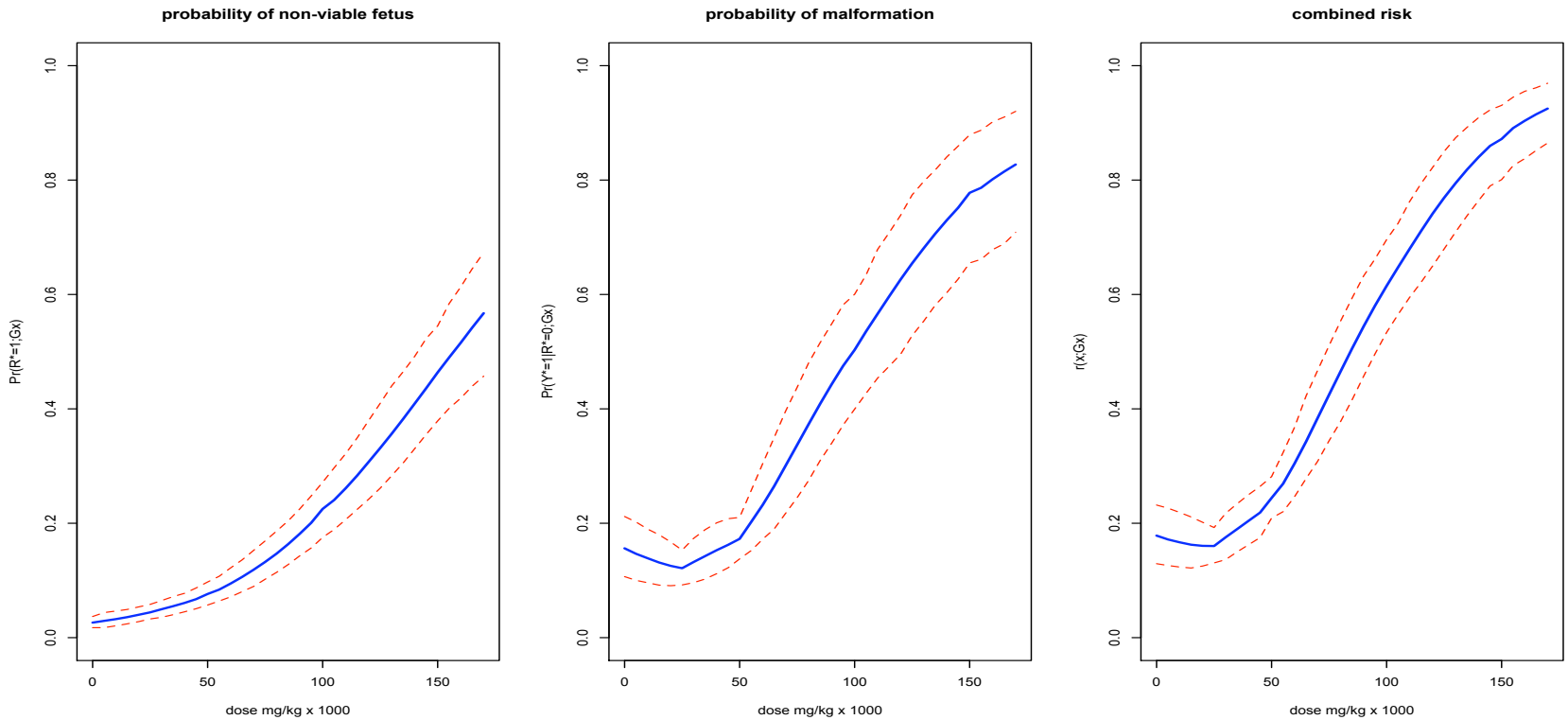


Figure 3.10: DEHP data. Posterior mean estimates and 90% uncertainty bands for the three dose-response curves. The model identifies the malformation endpoint as the sole contributor to the hormetic shape of the combined risk function.