Bayesian Nonparametric Mixture Modelling: Methods and Applications

Presenter: Athanasios Kottas
Department of Applied Mathematics and Statistics
University of California, Santa Cruz
(thanos@ams.ucsc.edu)

Co-presenter: Milovan Krnjajić
School of Mathematics, Statistics and Applied Mathematics
National University of Ireland, Galway
(milovan.krnjajic@nuigalway.ie)

National University of Ireland, Galway
December 14–15, 2009

Course Topics

- Notes 1: Dirichlet process priors (definitions, properties, and applications); Other nonparametric priors
- Notes 2: Dirichlet process mixture models – Methodology (definitions, examples, posterior simulation methods)
- Notes 3: Dirichlet process mixture models – Applications
- Notes 4: Dependent Dirichlet process models

Notes 1: Dirichlet process priors (definitions, properties, and applications); Other nonparametric priors

Outline

1.1 Bayesian nonparametrics

- An oxymoron?
- Priors on spaces of functions, $\{g(\cdot) : g \in \mathcal{G}\}$, vs usual parametric priors on Θ, where $g(\cdot) \equiv g(\cdot; \theta)$, $\theta \in \Theta$
- In certain applications, we may seek restrictions on the class of functions, e.g., monotone regression functions or unimodal error densities
- Functions of a univariate argument: distribution or density function, hazard or cumulative hazard function, link function, calibration function ...
- More generally, enriching usual parametric models, typically leading to semiparametric models
- Wandering nonparametrically near a standard class

Bayesian nonparametrics

- What objects are we modeling?
- A frequent goal is **means** (*Nonparametric Regression*)
- Usual approach: $g(x;\theta) = \sum_{k=1}^{K} \theta_k h_k(x)$
  where $\{h_k(x) : k = 1, ..., K\}$ is a collection of basis functions (splines, wavelets, Fourier series ...) – very large literature here
- An alternative is to use process realizations, i.e., $\{g(x) : x \in \mathcal{X}\}$, e.g., $g(\cdot)$ is a realization from a Gaussian process over $\mathcal{X}$

Bayesian nonparametrics

- Main focus: Modeling **random distributions**
- Distributions can be over scalars, vectors, even over a stochastic process
- Parametric modeling: based on parametric families of distributions $\{G(\cdot;\theta) : \theta \in \Theta\}$ – requires prior distributions over Θ
- Seek a richer class, i.e., $\{G : G \in \mathcal{G}\}$ – requires *nonparametric* prior distributions over $\mathcal{G}$
- How to choose $\mathcal{G}$? – how to specify the prior over $\mathcal{G}$? – requires specifying prior distributions for infinite-dimensional parameters

Bayesian nonparametrics

- What makes a nonparametric model "good"? (e.g., Ferguson, 1973)
  - The model should be tractable, in particular, it should yield inference that is readily available, either analytically or through simulations
  - The model should be rich, in the sense of having *large support*
  - The model hyperparameters should be easily interpretable

Bayesian nonparametrics

- General review papers on Bayesian nonparametrics: Walker, Damien, Laud & Smith (1999); Müller & Quintana (2004); Hanson, Branscum & Johnson (2005)

- Review papers on specific application areas of Bayesian nonparametric and semiparametric methods: Hjort (1996); Sinha & Dey (1997); Gelfand (1999)

- Books: Dey, Müller & Sinha (1998) (edited volume with a collection of papers, mainly, on applications of Bayesian nonparametrics); Ghosh & Ramamoorthi (2003) (emphasis on theoretical development of Bayesian nonparametric priors)

1.2 The Dirichlet process

- A Bayesian nonparametric approach to modeling, say, distribution functions requires priors for spaces of distribution functions
- Formally, it requires stochastic processes with sample paths that are distribution functions defined on an appropriate sample space $\mathcal{X}$ (e.g., $\mathcal{X} = R$, or R^+, or R^d), equipped with a σ-field $\mathcal{B}$ of subsets of $\mathcal{X}$ (e.g., the Borel σ-field for $\mathcal{X} \subseteq R^d$)
- The **Dirichlet process** (DP), anticipated in the work of Freedman (1963) and Fabius (1964), and formally developed by Ferguson (1973, 1974), is the first prior defined for spaces of distribution functions
- The DP is, formally, a (random) probability measure on the space of probability measures (distributions) on $(\mathcal{X}, \mathcal{B})$
- Hence, the DP generates random distributions on $(\mathcal{X}, \mathcal{B})$, and thus, for $\mathcal{X} \subseteq R^d$, equivalently, random cdfs on $\mathcal{X}$

The Dirichlet process

- Suppose you are dealing with a sample space with only two outcomes, say, $\mathcal{X} = \{0, 1\}$ and you are interested in estimating x, the probability of observing 1
- A natural prior for x is a beta distribution,

$$p(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} \qquad 0 \leq x \leq 1$$

- More generally, if $\mathcal{X}$ is finite with q elements, the probability distribution over $\mathcal{X}$ is given by q numbers $x_1, \ldots, x_q$ such that $\sum_{i=1}^{q} x_i = 1$. A natural prior for $(x_1, \ldots, x_q)$, which generalizes the Beta distribution, is the Dirichlet distribution (see next slide)
- With the Dirichlet process we further generalize to infinite but countable spaces

The Dirichlet process

Recall some properties of the Dirichlet distribution

- Start with independent rvs $Z_j \sim \text{gamma}(a_j, 1)$, $j = 1, ..., k$ (with $a_j > 0$)
- Define $Y_j = Z_j / (\sum_{\ell=1}^{k} Z_\ell)$, for $j = 1, ..., k$
- Then $(Y_1, ..., Y_k) \sim \text{Dirichlet}(a_1, ..., a_k)$ (distribution supported on R^{k-1}, since $\sum_{j=1}^{k} Y_j = 1$)
- $(Y_1, ..., Y_{k-1})$ has density $C(1 - \sum_{j=1}^{k-1} y_j)^{a_k - 1} \prod_{j=1}^{k-1} y_j^{a_j - 1}$, where $C = \Gamma(\sum_{j=1}^{k} a_j) / \{\prod_{j=1}^{k} \Gamma(a_j)\}$
- Moments: $\text{E}(Y_j) = a_j / \sum_{\ell=1}^{k} a_\ell$, $\text{E}(Y_j^2) = a_j(a_j + 1) / \{\sum_{\ell=1}^{k} a_\ell (1 + \sum_{\ell=1}^{k} a_\ell)\}$, and, for $i \neq j$, $\text{E}(Y_i Y_j) = a_i a_j / \{\sum_{\ell=1}^{k} a_\ell (1 + \sum_{\ell=1}^{k} a_\ell)\}$
- Note that for $k = 2$, $\text{Dirichlet}(a_1, a_2) \equiv \text{Beta}(a_1, a_2)$

The Dirichlet process

- The DP is characterized by two parameters:
  $\rightarrow$ Q_0 a specified probability measure on $(\mathcal{X}, \mathcal{B})$ (equivalently, G_0 a specified distribution function on $\mathcal{X}$)
  $\rightarrow$ α a positive scalar parameter

- **DEFINITION** (Ferguson, 1973): The DP generates random probability measures (random distributions) Q on $(\mathcal{X}, \mathcal{B})$ such that for any finite measurable partition B_1,...,B_k of $\mathcal{X}$,

$$(Q(B_1), ..., Q(B_k)) \sim \text{Dirichlet}(\alpha Q_0(B_1), ..., \alpha Q_0(B_k))$$

  $\rightarrow$ here, $Q(B_i)$ (a random variable) and $Q_0(B_i)$ (a constant) denote the probability of set B_i under Q and Q_0, respectively
  $\rightarrow$ also, the B_i, $i = 1, ..., k$, define a measurable partition if $B_i \in \mathcal{B}$, they are pairwise disjoint, and their union is $\mathcal{X}$

The Dirichlet process

- For any measurable subset B of $\mathcal{X}$, we have from the definition that $Q(B) \sim \text{Beta}(\alpha Q_0(B), \alpha Q_0(B^c))$, and thus

$$\text{E}(Q(B)) = Q_0(B)$$

and

$$\text{Var}(Q(B)) = \frac{Q_0(B)\{1 - Q_0(B)\}}{\alpha + 1},$$

- Q_0 plays the role of the *center* of the DP (also referred to as base probability measure, or base distribution)
- α can be viewed as a precision parameter: for large α there is small variability in DP realizations; the larger α is, the *closer* we expect a realization Q from the process to be to Q_0

- See Ferguson (1973) for the role of Q_0 on more technical properties of the DP (e.g., Ferguson shows that the support of the DP contains all probability measures on $(\mathcal{X}, \mathcal{B})$ that are absolutely continuous w.r.t. Q_0)

The Dirichlet process

- Analogously, for the random distribution function G on $\mathcal{X}$ generated from a DP with parameters α and G_0, a specified distribution function on $\mathcal{X}$
- For example, with $\mathcal{X} = R$, $B = (-\infty, x]$, $x \in R$, and $Q(B) = G(x)$,

$$G(x) \sim \text{Beta}(\alpha G_0(x), \alpha\{1 - G_0(x)\})$$

and thus

$$\text{E}(G(x)) = G_0(x)$$

and

$$\text{Var}(G(x)) = \frac{G_0(x)\{1 - G_0(x)\}}{\alpha + 1}$$

- **notation**: depending on the context, G will denote either the random distribution (probability measure) or the random distribution function $G \sim \text{DP}(\alpha, G_0)$ will indicate that a DP prior is placed on G

The Dirichlet process

- The definition can be used to simulate sample paths (which are distribution functions) from the DP — this is convenient when $\mathcal{X} \subseteq R$
- Consider any grid of points $x_1 < x_2 < ... < x_k$ in $\mathcal{X}$
- Then, the random vector $(G(x_1), G(x_2) - G(x_1), ..., G(x_k) - G(x_{k-1}), 1 - G(x_k))$ follows a Dirichlet distribution with parameters $(\alpha G_0(x_1), \alpha(G_0(x_2) - G_0(x_1)), ..., \alpha(G_0(x_k) - G_0(x_{k-1})), \alpha(1 - G_0(x_k)))$
- Hence, if $(u_1, u_2, ..., u_k)$ is a draw from this Dirichlet distribution, then $(u_1, ..., \sum_{j=1}^{i} u_j, ..., \sum_{j=1}^{k} u_j)$ is a draw from the distribution of $(G(x_1), ..., G(x_i), ..., G(x_k))$
- Example (Figure 1.1): $\mathcal{X} = (0, 1)$, $G_0(x) = x$, $x \in (0, 1)$ (Unif$(0, 1)$ base distribution)

The Dirichlet process

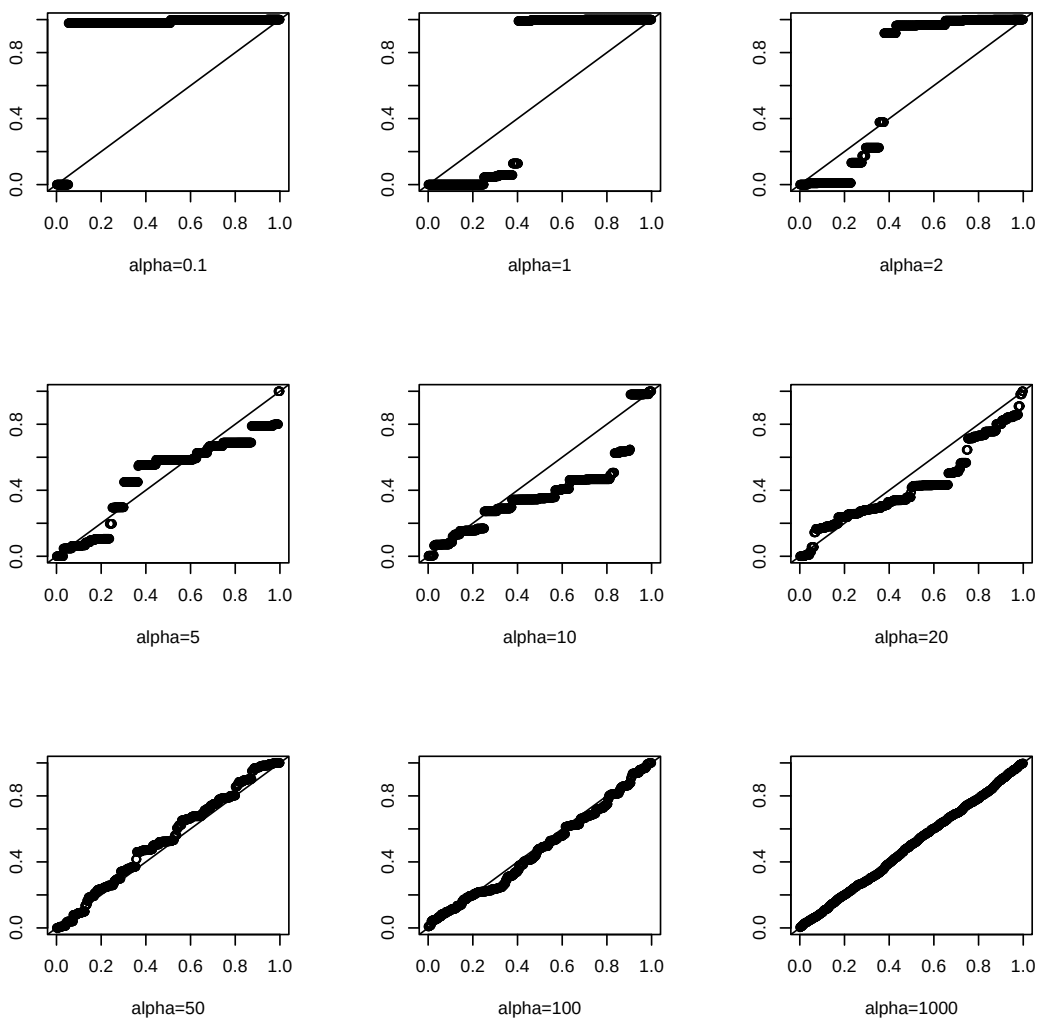


Figure 1.1: Cdf sample paths from a DP($\alpha, G_0 = \text{Unif}(0,1)$) prior, for different values of α. The solid line denotes the cdf of G_0.

The Dirichlet process

- **Constructive definition of the DP**

  (Sethuraman & Tiwari, 1982; Sethuraman, 1994)

  $\rightarrow$ let $\{z_r : r = 1, 2, ...\}$ and $\{\vartheta_\ell : \ell = 1, 2, ...\}$ be independent sequences of i.i.d. random variables:

  $*$ $z_r \sim \text{Beta}(1, \alpha)$, $r = 1, 2, ...$

  $*$ $\vartheta_\ell \sim G_0$, $\ell = 1, 2, ...$

  $\rightarrow$ define $\omega_1 = z_1$, $\omega_\ell = z_\ell \prod_{r=1}^{\ell-1}(1 - z_r)$, $\ell = 2, 3, ...$ (thus, $\sum_{\ell=1}^{\infty} \omega_\ell = 1$)

  $\rightarrow$ then, a realization G from $\text{DP}(\alpha, G_0)$ is (almost surely) of the form

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

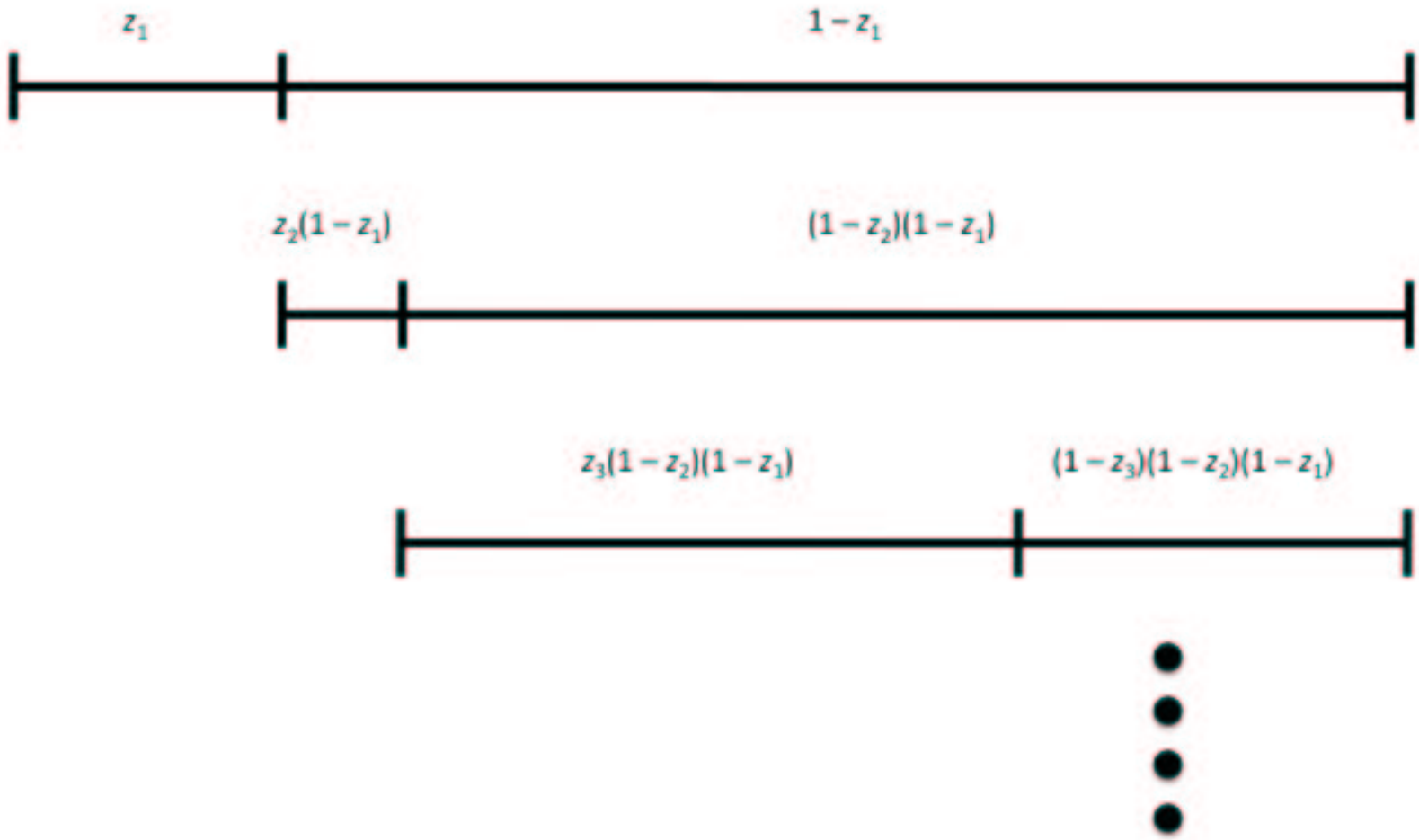
  (here, $\delta_z(\cdot)$ denotes a point mass at z)

- Hence, the DP generates distributions that have an (almost sure) representation as countable mixtures of point masses — the locations ϑ_ℓ are i.i.d. draws from the base distribution — their associated weights ω_ℓ are defined using the *stick-breaking* construction above

The Dirichlet process

- More on the DP stick-breaking construction:
  - Start with a stick of length 1 (representing the total probability to be distributed among the different atoms)
  - Draw a random $z_1 \sim \text{Beta}(1, \alpha)$, which defines the portion of the original stick assigned to atom 1, so that $\omega_1 = z_1$ — then, the remaining part of the stick has length $1 - z_1$
  - Draw a random $z_2 \sim \text{Beta}(1, \alpha)$ (independently of z_1), which defines the portion of the remaining stick assigned to atom 2, therefore, $\omega_2 = z_2(1 - z_1)$ — now, the remaining part of the stick has length $(1 - z_2)(1 - z_1)$
  - Continue ad infinitum ....
- We denote the joint distribution on the vector of weights by $(\omega_1, \omega_2, \dots) \sim SB(\alpha)$

The Dirichlet process



The Dirichlet process

- Based on its constructive definition, it is evident that the DP generates (almost surely) discrete distributions on $\mathcal{X}$ (this result was proved, using different approaches, by Ferguson, 1973, and Blackwell, 1973)
- The DP constructive definition yields another method to simulate from DP priors — in fact, it provides (up to a truncation approximation) the entire distribution G, not just cdf sample paths — for example, a possible approximation is

$$G_J = \sum_{j=1}^{J} p_j \delta_{\vartheta_j},$$

with $p_j = \omega_j$, $j = 1, ..., J-1$, and $p_J = 1 - \sum_{j=1}^{J-1} \omega_j = \prod_{r=1}^{J-1}(1 - z_r)$
$\rightarrow$ to specify J, note, for example, that

$$\mathrm{E}(\textstyle\sum_{j=1}^{J} \omega_j) = \mathrm{E}(1 - \prod_{r=1}^{J}(1 - z_r)) = 1 - \prod_{r=1}^{J} \mathrm{E}(1 - z_r) = 1 - \prod_{r=1}^{J} \frac{\alpha}{\alpha + 1} = 1 - (\frac{\alpha}{\alpha + 1})^J$$

$\rightarrow$ hence, J can be chosen such that $(\alpha/(\alpha+1))^J = \varepsilon$, for small ε

The Dirichlet process

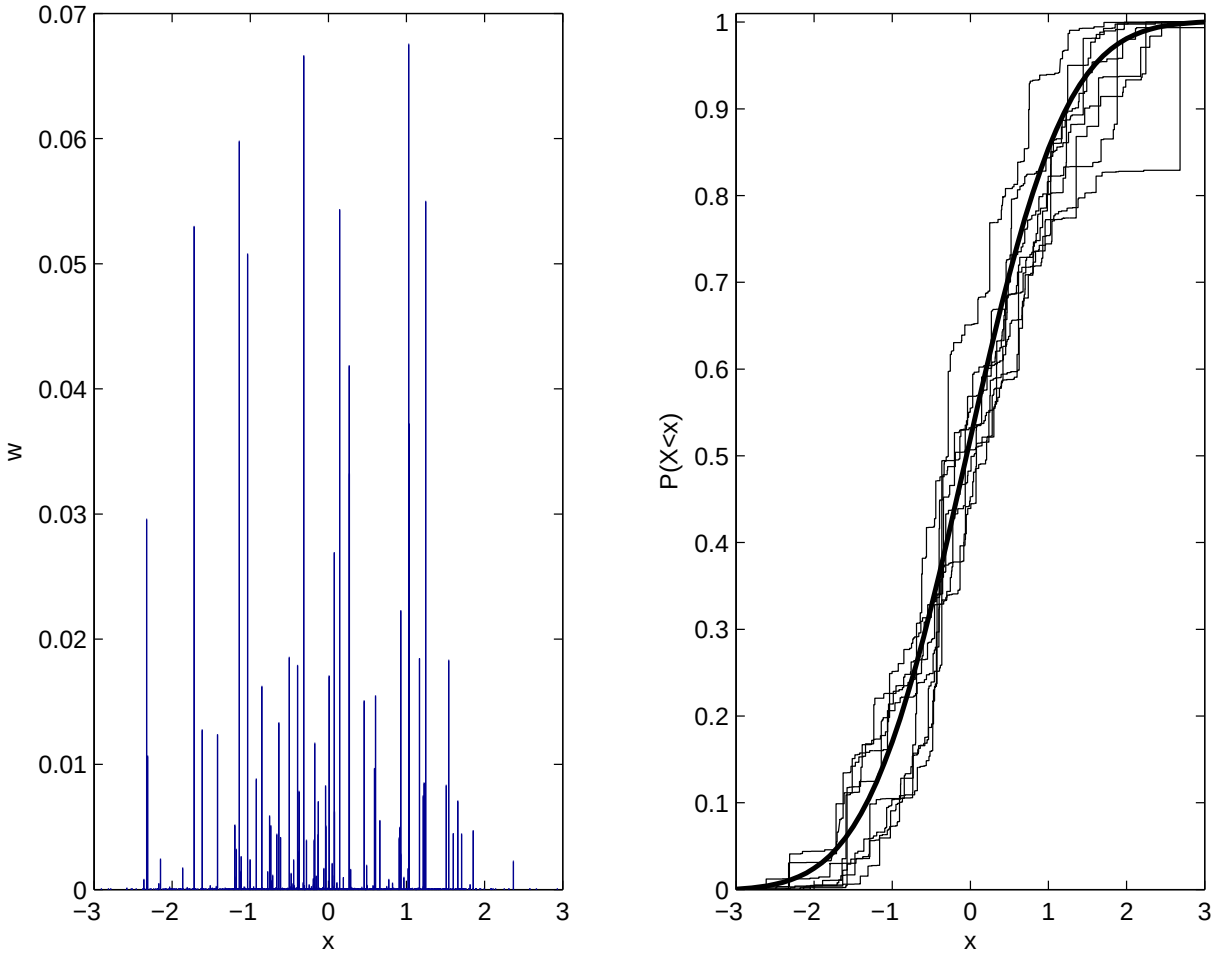


Figure 1.2: Illustration for a DP with $G_0 = N(0, 1)$ and $\alpha = 20$. In the left panel, the spiked lines are located at 1000 sampled values of x drawn from $N(0, 1)$ with heights given by the weights, ω_ℓ, calculated using the stick-breaking algorithm (a truncated version so that the weights sum to 1). These spikes are then summed from left to right to generate one cdf sample path from the DP. The right panel shows 8 such sample paths indicated by the lighter jagged lines. The heavy smooth line indicates the $N(0, 1)$ cdf.

- Moreover, the constructive definition of the DP has motivated several of its extensions, including:
  → the ϵ-DP (Muliere & Tardella, 1998); generalized DPs (Hjort, 2000); general stick-breaking priors (Ishwaran & James, 2001)
  → dependent DP priors (MacEachern, 1999, 2000; De Iorio et al., 2004; Griffin & Steel, 2006)
  → hierarchical DPs (Tomlinson & Escobar, 1999; Teh et al., 2006)
  → spatial DP models (Gelfand, Kottas & MacEachern, 2005; Kottas, Duan & Gelfand, 2008; Duan, Guindani & Gelfand, 2007)
  → nested DPs (Rodriguez, Dunson & Gelfand, 2008)

- **Pólya urn characterization of the DP**

  (Blackwell & MacQueen, 1973)

  → if, for $i = 1, ..., n$, $x_i \mid G$ are i.i.d. from G, and $G \sim \text{DP}(\alpha, G_0)$, then, marginalizing G over its DP prior, the induced joint distribution for the x_i is given by

$$p(x_1, ..., x_n) = G_0(x_1) \prod_{i=2}^{n} \left\{ \frac{\alpha}{\alpha + i - 1} G_0(x_i) + \frac{1}{\alpha + i - 1} \sum_{j=1}^{i-1} \delta_{x_j}(x_i) \right\}$$

  → that is, the sequence of the x_i follows a generalized Pólya urn scheme such that

  $*$ $x_1 \sim G_0$, and

  $*$ for any $i = 2, ..., n$, $x_i \mid x_1, ..., x_{i-1}$ follows the mixed distribution that places point mass $(\alpha + i - 1)^{-1}$ at x_j, $j = 1, ..., i - 1$, and continuous mass $\alpha(\alpha + i - 1)^{-1}$ on G_0

- **Prior to posterior updating with DP priors**

  (Ferguson, 1973)

  $\rightarrow$ let G denote the random distribution function for the following results

  $\rightarrow$ if the observations $y_i \mid G$ are i.i.d. from G, $i = 1, ..., n$, and $G \sim \text{DP}(\alpha, G_0)$, then the posterior distribution of G is a $\text{DP}(\tilde{\alpha}, \tilde{G}_0)$, with $\tilde{\alpha} = \alpha + n$, and

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

- Hence, the DP is a *conjugate* prior — all the results and properties developed for DPs can be used directly for the posterior distribution of G

- For example, the posterior point estimate for $G(t)$

$$\mathrm{E}(G(t) \mid y_1, ..., y_n) = \frac{\alpha}{\alpha + n} G_0(t) + \frac{n}{\alpha + n} G_n(t)$$

where $G_n(t) = n^{-1} \sum_{i=1}^{n} 1_{[y_i, \infty)}(t)$ is the empirical distribution function of the data (the standard classical nonparametric estimator)
$\rightarrow$ for small α relative to n, little weight is placed on the prior guess G_0
$\rightarrow$ for large α relative to n, little weight is placed on the data
$\rightarrow$ α can be viewed as a measure of faith in the prior guess G_0 measured in units of number of observations (thus, $\alpha = 1$ indicates strength of belief in G_0 worth one observation)

- **Mixtures of Dirichlet processes** (Antoniak, 1974)
Extension of the DP to a hierarchical version: $G \mid \alpha, \psi \sim \mathrm{DP}(\alpha, G_0(\cdot \mid \psi))$, where (parametric) priors are added to the precision parameter α and/or the parameters, ψ, of the base distribution

Generalizing the DP

- Many random probability measures can be defined by means of a stick-breaking construction – the z_r are drawn independently from a distribution on $[0, 1]$
- For example, the Beta two-parameter process (Ishwaran & Zarepour, 2000) is defined by choosing $z_r \sim \text{Beta}(a, b)$
- If $z_r \sim \text{Beta}(1 - a, b + ra)$, $r = 1, 2, \ldots$, for some $a \in [0, 1)$ and $b \in (-a, \infty)$, we obtain the two-parameter Poisson-Dirichlet process (e.g., Pitman & Yor, 1997)
- The general case, $z_r \sim \text{Beta}(a_r, b_r)$ (Ishwaran & James, 2001)

- More generally, Ongaro and Cattaneo (2004) consider the discrete random probability measure

$$G_K(\cdot) = \sum_{k=1}^{K} p_k \, \delta_{\theta_k^*}(\cdot),$$

where K is an integer random variable (allowed to be infinite); and conditionally on K, the θ_k^* are i.i.d. from some base distribution G_0 (not necessarily nonatomic), and the weights p_k are allowed to have any distribution on the simplex

$$\{\boldsymbol{p} : \sum_{k=1}^{K} p_k = 1; p_k \geq 0, k = 1, \ldots, K\}$$

1.3 Dose-response modeling with Dirichlet process priors

- **Quantal bioassay problem**: study potency of a stimulus by administering it at k dose levels to a number of subjects at each level
  $\rightarrow x_i$: dose levels (with $x_1 < x_2 < ... < x_k$)
  $\rightarrow n_i$: number of subjects at dose level i
  $\rightarrow y_i$: number of positive responses at dose level i

- $F(x) = \Pr(\text{positive response at dose level } x)$ (i.e., the *potency* of level x of the stimulus) — F is referred to as the potency curve, or dose-response curve, or tolerance distribution

- Standard assumption in bioassay settings: the probability of a positive response increases with increasing dose level, i.e., F is a non-decreasing function, i.e., F can be modeled as a cdf on $\mathcal{X} \subseteq R$

Dose-response modeling with Dirichlet process priors

- Parametric modeling: F is assumed to be a member of a parametric family of cdfs (e.g., logit, or probit models)
- Bayesian nonparametric modeling: uses a nonparametric prior for the infinite dimensional parameter F, i.e., a prior for the space of cdfs on $\mathcal{X}$ — work based on a DP prior for F: Antoniak (1974), Bhattacharya (1981), Disch (1981), Kuo (1983, 1988), Gelfand & Kuo (1991), Mukhopadhyay (2000)

- Questions of interest:
  1. Inference for $F(x)$ for specified dose levels x
  2. Inference for unobserved dose level x_0 such that $F(x_0) = \gamma$ for specified $\gamma \in (0, 1)$
  3. Optimal selection of $\{x_i, n_i\}$ to best accomplish goals 1 and 2 above (design problem)

Dose-response modeling with Dirichlet process priors

- Assuming independent outcomes at different dose levels, the likelihood is given by $\prod_{i=1}^{k} p_i^{y_i}(1-p_i)^{n_i-y_i}$, where $p_i = F(x_i)$, $i = 1, ..., k$
- If the prior for F is a DP with precision parameter $\alpha > 0$ and base cdf F_0 (the prior guess for the potency curve), the induced prior on $(p_1, ..., p_k)$ is an ordered Dirichlet distribution, i.e.,
  $(p_1, p_2 - p_1, ..., p_k - p_{k-1}, 1 - p_k)$ follows a Dirichlet distribution with parameters
  $(\alpha F_0(x_1), \alpha(F_0(x_2) - F_0(x_1)), ..., \alpha(F_0(x_k) - F_0(x_{k-1})), \alpha(1 - F_0(x_k)))$

- The posterior for F is a mixture of Dirichlet processes (Antoniak, 1974)
  $\rightarrow$ posterior distribution is difficult to work with analytically (Antoniak obtained point estimate when $k = 2$)
  $\rightarrow$ Markov chain Monte Carlo (MCMC) techniques enable full inference (e.g., Gelfand & Kuo, 1991; Mukhopadhyay, 2000)

1.4 Bayesian nonparametric modeling for cytogenetic dosimetry

- Cytogenetic dosimetry (in vitro setting): samples of cell cultures exposed to a range of doses of a given agent — in each sample, at each dose level, a measure of cell disability is recorded
- Dose-response modeling framework, where "dose" is the form of exposure to radiation, and "response" is the measure of genetic aberration (in vivo setting, human exposures), or cell disability (in vitro setting, cell cultures of human lymphocytes)
- Focus on categorical classification for the response
  → binary response (1 positive response, 0 no response) — bioassay problem
  → (ordered) polytomous response (requires priors on two or more functions)

Bayesian nonparametric modeling for cytogenetic dosimetry

- For polytomous responses:
  $\rightarrow x_i$: dose levels (with $x_1 < x_2 < ... < x_k$)
  $\rightarrow n_i$: number of cells at dose level i
  $\rightarrow \boldsymbol{y}_i = (y_{i1}, ..., y_{ir})$: response vector ($r \geq 2$ classifications) at dose level i
- Hence, now $\boldsymbol{y}_i \sim \text{Mult}(n_i, \boldsymbol{p}_i)$, where $\boldsymbol{p}_i = (p_{i1}, ..., p_{ir})$

- Data set (Madruga et al., 1996): blood samples from individuals exposed in vitro to ^{60}Co radiation with doses 20, 50, 100, 200, 300, 400, and 500 centograms — lymphocyte cultures prepared for a cytokinesis-block micronucleus assay — response: presence of binucleated cells with 0, 1, or ≥ 2 micronuclei — use of these $r = 3$ classifications rather than the actual counts arises because, when there are multiple micronuclei, it is difficult for the assayers to count the exact number
- Questions of interest: 1. Prediction of response at "new" dose levels
  2. Inference for unknown doses (exposures) given observed responses (this inversion problem is practically important, since although the response is typically accurately observed, the exposure is difficult to measure)

Bayesian nonparametric modeling for cytogenetic dosimetry

- Bayesian nonparametric modeling for polytomous response (Kottas, Branco & Gelfand, 2002)
- Consider simple case with $r = 3$ — model for p_{i1} and p_{i2} is needed
- Model $p_{i1} = F_1(x_i)$ and $p_{i1} + p_{i2} = F_2(x_i)$, and thus $F_1(\cdot) \leq F_2(\cdot)$
- Bayesian nonparametric model requires prior on the space

$$\{(F_1, F_2) : F_1(\cdot) \leq F_2(\cdot)\}$$

  of stochastically ordered pairs of cdfs (F_1, F_2)
- Constructive approach: $F_1(\cdot) = G_1(\cdot)G_2(\cdot)$, and $F_2(\cdot) = G_1(\cdot)$ with independent DP$(\alpha_\ell, G_{0\ell})$ priors for G_ℓ, $\ell = 1, 2$
- Induced prior for $\boldsymbol{q}_\ell = (q_{\ell,1}, ..., q_{\ell,k})$, $\ell = 1, 2$, where $q_{\ell,i} = G_\ell(x_i)$

- Combining with the likelihood, the posterior for $(\boldsymbol{q}_1, \boldsymbol{q}_2)$ is given by

$$\begin{aligned} p(\boldsymbol{q}_1, \boldsymbol{q}_2 \mid \text{data}) \quad &\propto \prod_{i=1}^{k} \left\{ q_{1i}^{y_{i1}+y_{i2}} (1-q_{1i})^{y_{i3}} q_{2i}^{y_{i1}} (1-q_{2i})^{y_{i2}} \right\} \\ &\times q_{11}^{\gamma_1 - 1} (q_{12}-q_{11})^{\gamma_2 - 1} \ldots (q_{1k}-q_{1,k-1})^{\gamma_k - 1} (1-q_{1k})^{\gamma_{k+1}-1} \\ &\times q_{21}^{\delta_1 - 1} (q_{22}-q_{21})^{\delta_2 - 1} \ldots (q_{2k}-q_{2,k-1})^{\delta_k - 1} (1-q_{2k})^{\delta_{k+1}-1} \end{aligned}$$

  where $\gamma_i = \alpha_1(G_{01}(x_i) - G_{01}(x_{i-1}))$ and $\delta_i = \alpha_2(G_{02}(x_i) - G_{02}(x_{i-1}))$

- Simulation-based model fitting yields posterior draws from $p(\boldsymbol{q}_1, \boldsymbol{q}_2 \mid \text{data})$

- Posteriors for $G_\ell(x_i)$, $\ell = 1, 2$, provide posteriors for $F_1(x_i)$ and $F_2(x_i)$, for all x_i, $i = 1, ..., k$

- For any unobserved dose level x_0, the posterior (predictive) distribution for $q_{\ell,0} = G_\ell(x_0)$, $\ell = 1, 2$, is given by

$$p(q_{\ell,0} \mid \text{data}) = \int p(q_{\ell,0} \mid \boldsymbol{q}_\ell) p(\boldsymbol{q}_\ell \mid \text{data}) \mathrm{d}\boldsymbol{q}_\ell$$

  where $p(q_{\ell,0} \mid \boldsymbol{q}_\ell)$ is a rescaled Beta distribution

- Hence, we can obtain the posterior for $G_\ell(x_0)$, $\ell = 1, 2$, for any set of x_0 values, and thus, we can obtain the posterior for $F_1(x_0)$ and $F_2(x_0)$ at any x_0 — yields posterior point and interval estimates for $F_1(\cdot)$ and $F_2(\cdot)$
- The inversion problem can also be handled: inference for unknown x_0 for specified values of $\boldsymbol{y}_0 = (y_{01}, y_{02}, y_{03})$ — extend the MCMC method to the augmented posterior that includes the additional parameter vector (x_0, q_{10}, q_{20})
- For the data illustrations, we compare with a parametric logit model

$$\log \frac{p_{ij}}{p_{i3}} = \beta_{1j} + \beta_{2j} x_i, \quad i = 1, ..., k, \quad j = 1, 2$$

(model fitting, prediction, and inversion are straightforward under this model)

Data Illustrations

Table 1.1: (Data from Madruga et al., 1996). 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. 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., $\widehat{\eta}_{i2} = (y_{i1} + y_{i2})/(y_{i1} + y_{i2} + y_{i3})$.

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

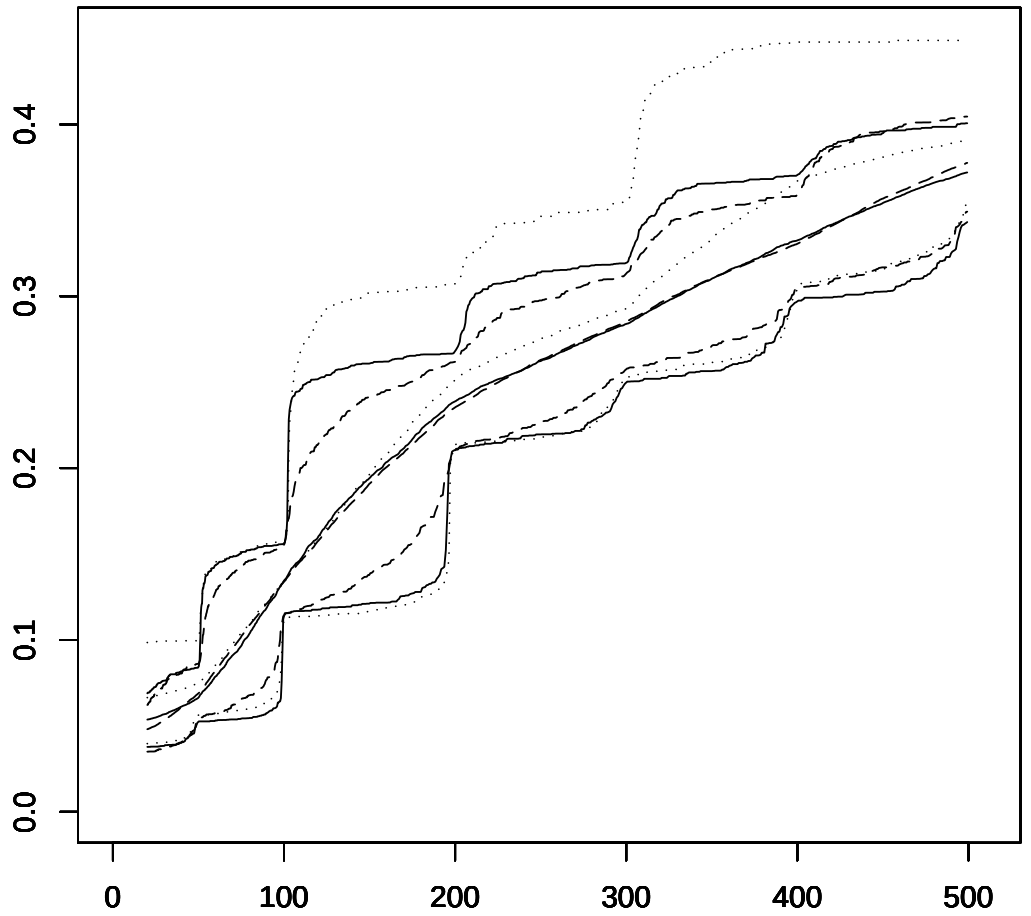


Figure 1.3: For the data in Table 1.1, point and 95% pointwise interval posterior estimates for the probability of at least one MN vs dose under $\alpha_1 = \alpha_2 = 0.1$ (dotted lines), 1 (solid lines) and 10 (dashed lines).

- Simulated data to compare the parametric and nonparametric models
- $r = 3$, $k = 7$, same dose values with the real data
- Two sample sizes: one with n_i as in Table 1, and one with smaller sample sizes, $n_i/10$

- Simulation 1: data generated from the parametric model
- Simulation 2: non-standard (bimodal) shapes for F_1 and F_2

Bayesian nonparametric modelling for cytogenetic dosimetry

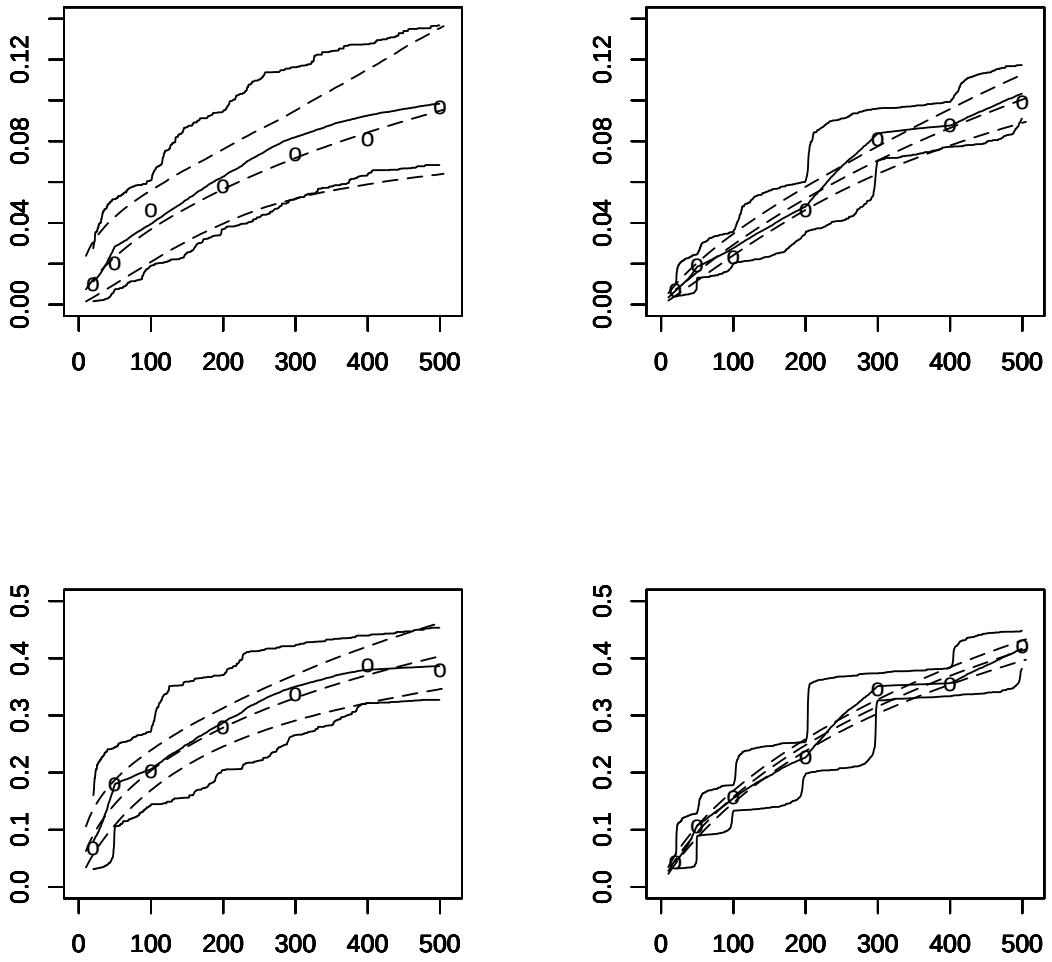


Figure 1.4: Simulation 1. Posterior inference for F_1 (upper panels) and F_2 (lower panels) under the parametric (dashed lines) and nonparametric (solid lines) model. "o" denotes the observed data. The left and right panels correspond to the data set with the smaller and large sample sizes, respectively.

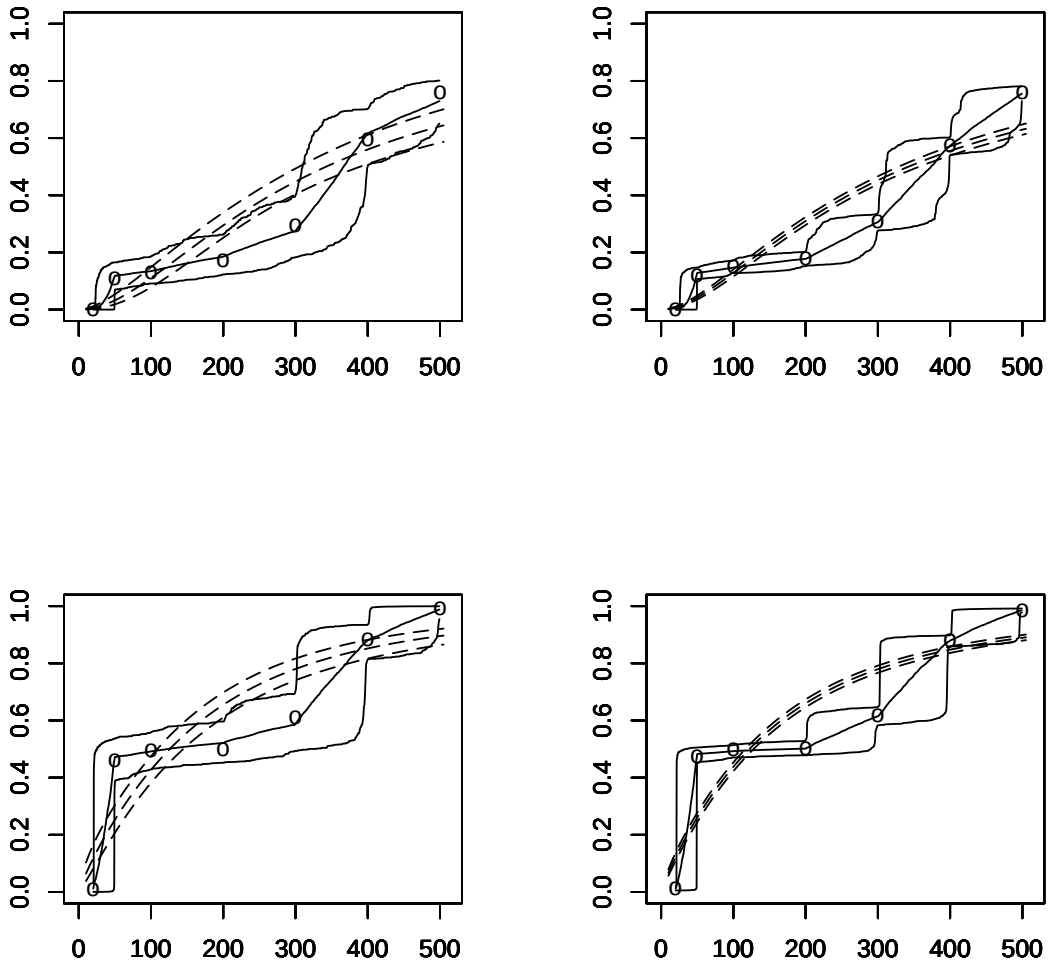


Figure 1.5: Simulation 2. Posterior inference for F_1 (upper panels) and F_2 (lower panels) under the parametric (dashed lines) and nonparametric (solid lines) model. "o" denotes the observed data. The left and right panels correspond to the data set with the smaller and large sample sizes, respectively.

1.5 Semiparametric regression for categorical responses

- Application of DP-based modeling to semiparametric regression with categorical responses
- Categorical responses y_i, $i = 1, ..., N$ (e.g., counts or proportions)
- Covariate vector $\boldsymbol{x}_i$ for the i-th response, comprising either categorical predictors or quantitative predictors with a finite set of possible values
- $K \leq N$ predictor profiles (cells), where each cell k $(k = 1, ..., K)$ is a combination of observed predictor values — $k(i)$ denotes the cell corresponding to the i-th response
- Assume that all responses in a cell are exchangeable with distribution F_k, $k = 1, ..., K$

Semiparametric regression for categorical responses

- *Product of mixtures of Dirichlet processes prior* (Cifarelli & Regazzini, 1978) for the cell-specific random distributions F_k, $k = 1, ..., K$
  $\rightarrow$ conditionally on hyperparameters α_k and $\boldsymbol{\theta}_k$, the F_k are assigned independent DP$(\alpha_k, F_{0k}(\cdot; \boldsymbol{\theta}_k))$ priors, where, in general, $\boldsymbol{\theta}_k = (\theta_{1k}, ..., \theta_{Dk})$
  $\rightarrow$ the F_k are related by modeling the α_k $(k = 1, ..., K)$ and/or the θ_{dk} $(k = 1, ..., K;\ d = 1, ..., D)$ as linear combinations of the predictors (through specified link functions h_d, $d = 0, 1, ..., D$)
  $\rightarrow$ $h_0(\alpha_k) = \boldsymbol{x}_k^T \boldsymbol{\gamma}$, $k = 1, ..., K$
  $\rightarrow$ $h_d(\theta_{dk}) = \boldsymbol{x}_k^T \boldsymbol{\beta}_d$, $k = 1, ..., K;\ d = 1, ..., D$
  $\rightarrow$ (parametric) priors for the vectors of regression coefficients $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}_d$

- DP-based prior model that induces dependence in the finite collection of distributions $\{F_1, ..., F_K\}$, though a weaker type of dependence than more recent approaches building on dependent DP priors (refer to the fourth set of notes)

Semiparametric regression for categorical responses

- Semiparametric structure centered around a *parametric backbone* defined by the $F_{0k}(\cdot;\boldsymbol{\theta}_k)$ — useful interpretation and connections with parametric regression models

- Example: regression model for counts (Carota & Parmigiani, 2002)

$$\begin{aligned} y_i \mid \{F_1, ..., F_K\} &\sim \prod_{i=1}^{N} F_{k(i)}(y_i) \\ F_k \mid \alpha_k, \theta_k &\overset{ind.}{\sim} \mathrm{DP}(\alpha_k, \mathrm{Poisson}(\cdot;\theta_k)), \quad k = 1, ..., K \\ \log(\alpha_k) = \boldsymbol{x}_k^T \boldsymbol{\gamma} &\qquad \log(\theta_k) = \boldsymbol{x}_k^T \boldsymbol{\beta}, \quad k = 1, ..., K \end{aligned}$$

  with priors for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$

- Related work for: change-point problems (Mira & Petrone, 1996); dose-response modeling for toxicology data (Dominici & Parmigiani, 2001); variable selection in survival analysis (Giudici, Mezzetti & Muliere, 2003)

1.6 Other Bayesian nonparametric approaches

Pólya tree priors

- Pólya tree processes (Ferguson, 1974; Mauldin, Sudderth & Williams, 1992; Lavine, 1992, 1994)
- Binary partitioning of the support for distribution G (so, most widely used on R^1): first, partition support into B_0, B_1; next, B_0 into B_{00}, B_{01}, and B_1 into B_{10}, B_{11}; etc.
- Finite Pólya trees: truncating at M levels (total of 2^M sets)
- Random probabilities assigned sequentially:
  $\rightarrow G(B_0) = \Pr(\theta \in B_0) = \nu_0$, where $\nu_0 \sim \text{Beta}(\alpha_0, \alpha_1)$
  $\rightarrow \Pr(\theta \in B_{00} | \theta \in B_0) = \nu_{00}$, with $\nu_{00} \sim \text{Beta}(\alpha_{00}, \alpha_{01})$, so $G(B_{00}) = \nu_0 \nu_{00}$
  $\rightarrow$ for example, $G(B_{1001}) = \nu_1 \nu_{10} \nu_{100} \nu_{1001}$
- The partition Π, determined by the collection of all the sets B, and the vector, $\mathcal{A}$, of all the α define a Pólya tree distribution $G|\Pi, \mathcal{A}$

Other Bayesian nonparametric approaches

- Centering around a specified distribution G_0?
- Define $B_0 = (-\infty, G_0^{-1}(0.5))$, $B_{00} = (-\infty, G_0^{-1}(0.25))$, etc.
- Set $\alpha_0 = \alpha_1$, set $\alpha_{00} = \alpha_{01}$, etc.
- Then $G(B_0) = \nu_0 \sim \text{Beta}(\alpha_0, \alpha_0)$, so $\text{E}(G(B_0)) = 0.5 = G_0(B_0)$
- Realizations of θ are in one of 2^M sets and can be represented through the dyadic partition, $\theta = G_0^{-1}(\sum_{j=1}^{M} \delta_j 2^{-j})$ (labeled by left endpoint)
  $\rightarrow$ for example, B_{1001} has $\delta_1 = 1$, $\delta_2 = 0$, $\delta_3 = 0$, $\delta_4 = 1$

- Conjugacy property: if $y_i|G$ are i.i.d. from G and G has a Pólya tree prior with specified parameters Π and $\mathcal{A}$, then the posterior of G, given the data y_i, is a Pólya tree distribution with updated parameters
- MCMC methods needed for models utilizing Pólya tree priors with random parameters Π and/or $\mathcal{A}$ (Hanson, 2006a) or Pólya trees with "jittered" partitions (Paddock et al., 2003)

Other Bayesian nonparametric approaches

- Choice of the α?
- The DP is a special case, i.e., $\alpha_{00} + \alpha_{01} = \alpha_0$, etc. (can verify that it produces the usual partitioning for the Dirichlet distribution)
- Consider c_m at level m, where $c_1 = \alpha_0 = \alpha_1$, $c_2 = \alpha_{00} = \alpha_{01} = \alpha_{10} = \alpha_{11}$, etc. – can argue that c_m should increase in m, e.g., $c_m = cm^2$ yields random process realizations that are (almost surely) continuous
- DP has $c_m = c/2^m$, i.e., the wrong direction with regard to continuity

- Modelling applications with Pólya tree priors: survival analysis (Muliere & Walker, 1997a; Walker & Mallick, 1999); bioassay modeling (Muliere & Walker, 1997b); median regression (Hanson & Johnson, 2002); multiple imputation with partially observed data (Paddock, 2002); ROC data analysis (Branscum et al., 2008; Hanson, Kottas & Branscum, 2008)

Stochastic process approach

- Usually applied to R^+ with suggestive argument t
- Write the random cdf F as $F(t) = 1 - e^{-Z(t)}$, where $Z(\cdot)$ is a *neutral to the right Lévy process* (e.g., Ferguson & Phadia, 1979) (i.e., $Z(\cdot)$ has independent increments, and is, almost surely, non-decreasing, right continuous, with $Z(0) = 0$ and $\lim_{t\to\infty} Z(t) = \infty$; in fact, $Z(\cdot)$ has, at most, countably many jumps)
- so, $Z(\cdot) = -\log(1 - F(\cdot))$ (modeling a cumulative hazard function rather than the cdf)
- A particular example is the Gamma process (e.g., Kalbfleisch, 1978)
  $\rightarrow$ Consider an arbitrary finite partition of R^+, $0 = a_0 < a_1 < a_2 ... < a_k < a_{k+1} = \infty$
  $\rightarrow$ Let $q_l = \Pr(T \in [a_{l-1}, a_l) | T \geq a_{l-1})$ and let $r_l = -\log(1 - q_l)$
  $\rightarrow$ then $\sum_{l=1}^{k} r_l = -\log \Pr(T \geq a_k) = Z(a_k)$

Other Bayesian nonparametric approaches

- So, think of $Z(\cdot)$ as a Gamma process: $Z(t_2) - Z(t_1) \sim \text{Gamma}(c(Z_0(t_2) - Z_0(t_1)), c)$, where $Z_0(\cdot)$ is a specified monotonic function and c is a precision constant
- r_l independent implies q_l independent (restrictive?)
- Incorporate covariates with $r_l(x) = r_l \exp(x^T \beta)$, i.e., a proportional hazards model, $\Pr(T \geq t) = \exp(-Z(t)) \exp(x^T \beta)$
- Connection with DP – under the DP prior, the q_l are i.i.d. Beta
- Cumulative hazard is a step function, so F is as well – steps can be erratic, smoothing?
- Alternatively, the hazard function can be modeled directly using the extended Gamma process (Dykstra & Laud, 1981)

Notes 2: Dirichlet process mixture models – Methodology (definitions, examples, posterior simulation methods)

Outline

2.1 Mixture distributions

2.2 Model details, examples, hierarchical formulation

2.3 Prior specification

2.4 Methods for posterior inference

2.1 Mixture distributions

- Mixture models arise naturally as flexible alternatives to standard parametric families
- Continuous mixture models (e.g., t, Beta-binomial and Poisson-gamma models) typically achieve increased heterogeneity but are still limited to unimodality and usually symmetry
- Finite mixture distributions provide more flexible modelling, and are now feasible to implement due to advances in simulation-based model fitting (e.g., Richardson & Green, 1997; Stephens, 2000; Jasra, Holmes & Stephens, 2005)
- Rather than handling the very large number of parameters of finite mixture models with a large number of mixands, it may be easier to work with an infinite dimensional specification by assuming a random mixing distribution, which is not restricted to a specified parametric family

Mixture distributions

- Recall the structure of a finite mixture model with K components, for example, a mixture of $K = 2$ Gaussian densities:

$$y_i \mid w, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \overset{ind.}{\sim} w\mathrm{N}(y_i; \mu_1, \sigma_1^2) + (1-w)\mathrm{N}(y_i; \mu_2, \sigma_2^2),$$

  that is, observation y_i arises from a $\mathrm{N}(\mu_1, \sigma_1^2)$ distribution with probability w or from a $\mathrm{N}(\mu_2, \sigma_2^2)$ distribution with probability $1-w$ (independently for each $i = 1, ..., n$)

- In the Bayesian setting, we also set priors for the unknown parameters

$$(w, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \sim p(w, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$$

- The model can be rewritten in a few different ways. For example, we can introduce auxiliary random variables $L_1, \ldots, L_n$ such that $L_i = 1$ if y_i arises from the $\mathrm{N}(\mu_1, \sigma_1^2)$ component (component 1) and $L_i = 2$ if y_i is drawn from the $\mathrm{N}(\mu_2, \sigma_2^2)$ component (component 2). Then, the model can be written as

$$y_i \mid L_i, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \overset{ind.}{\sim} \mathrm{N}(y_i; \mu_{L_i}, \sigma_{L_i}^2)$$
$$\Pr(L_i = 1|w) = w = 1 - \Pr(L_i = 2|w)$$
$$(w, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \sim p(w, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$$

- If we marginalize over L_i we recover the original mixture formulation. The inclusion of indicator variables is very common in finite mixture models, and we will make extensive use of it

- We can also write

$$w\mathrm{N}(y_i;\mu_1,\sigma_1^2)+(1-w)\mathrm{N}(y_i;\mu_2,\sigma_2^2)=\int \mathrm{N}(y_i;\mu,\sigma^2)\mathrm{d}G(\mu,\sigma^2)$$

$$G(\cdot)=w\delta_{(\mu_1,\sigma_1^2)}(\cdot)+(1-w)\delta_{(\mu_2,\sigma_2^2)}(\cdot)$$

  A similar expression can be used for a general K mixture model

- Note that G is discrete (and random) — a natural alternative is to use a DP prior for G, resulting in a Dirichlet process mixture (DPM) model

- Working with a countable mixture (rather than a finite one) provides theoretical advantages (full support) as well as practical (the model *automatically* decides how many components are appropriate for a given data set). Recall the comment about "good" nonparametric/semiparametric Bayes

2.2 Model details, examples, hierarchical formulation

- The Dirichlet process (DP) has been the most widely used prior for the random mixing distribution, following the early work by Antoniak (1974), Lo (1984) and Ferguson (1983)

  Dirichlet process mixture model

$$F(\cdot; G) = \int K(\cdot; \theta) \mathrm{d}G(\theta), \qquad G \sim \mathrm{DP}(\alpha, G_0)$$

  with $K(\cdot; \theta)$ a parametric family of distribution functions indexed by θ

- Corresponding mixture density (or probability mass) function,

$$f(\cdot; G) = \int k(\cdot; \theta) \mathrm{d}G(\theta)$$

  where $k(\cdot; \theta)$ is the density (or probability mass) function of $K(\cdot; \theta)$

- Because G is random, the distribution function $F(\cdot; G)$ and the density function $f(\cdot; G)$ are random (Bayesian nonparametric mixture models)

- Contrary to DP prior models, the DP mixture $F(\cdot; G)$ can model both discrete distributions (e.g., $K(\cdot; \theta)$ might be Poisson or binomial) and continuous distributions, either univariate ($K(\cdot; \theta)$ can be, e.g., normal, gamma, or uniform) or multivariate (with $K(\cdot; \theta)$, say, multivariate normal)

- Several useful results for general mixtures of parametric families, e.g.,
  $\rightarrow$ (discrete) normal location-scale mixtures, $\sum_{j=1}^{M} w_j \mathrm{N}(\cdot \mid \mu_j, \sigma_j^2)$, can approximate arbitrarily well any density on the real line (Lo, 1984; Ferguson, 1983; Escobar & West, 1995) — analogously, for densities on R^d (West et al., 1994; Müller et al., 1996)
  $\rightarrow$ for any non-increasing density $f(t)$ on the positive real line there exists a distribution function G such that f can be represented as a scale mixture of uniform densities, i.e., $f(t) = \int \theta^{-1} 1_{[0,\theta)}(t) \mathrm{d}G(\theta)$ — the result yields flexible DP mixture models for symmetric unimodal densities (Brunner & Lo, 1989; Brunner, 1995) as well as general unimodal densities (Brunner, 1992; Lavine & Mockus, 1995; Kottas & Gelfand, 2001; Kottas & Krnjajić, 2009)

- Typically, semiparametric DP mixtures are employed

$$\begin{aligned} y_i \mid G, \phi &\overset{i.i.d.}{\sim} f(\cdot; G, \phi) = \int k(\cdot; \theta, \phi) \mathrm{d}G(\theta), \quad i = 1, ..., n \\ G &\sim \mathrm{DP}(\alpha, G_0) \end{aligned}$$

  with a parametric prior $p(\phi)$ placed on ϕ (and, perhaps, hyperpriors for α and/or the parameters ψ of $G_0 \equiv G_0(\cdot \mid \psi)$)

- **Hierarchical formulation** for DP mixture models: introduce latent mixing parameter θ_i associated with y_i

$$\begin{aligned} y_i \mid \theta_i, \phi &\overset{ind.}{\sim} k(y_i; \theta_i, \phi), \quad i = 1, ..., n \\ \theta_i \mid G &\overset{i.i.d.}{\sim} G, \quad i = 1, ..., n \\ G \mid \alpha, \psi &\sim \mathrm{DP}(\alpha, G_0), \;\; G_0 = G_0(\cdot \mid \psi) \\ \phi, \alpha, \psi &\sim p(\phi)p(\alpha)p(\psi) \end{aligned}$$

Model details, examples, hierarchical formulation

- In the context of DP mixtures, the (almost sure) discreteness of realizations G from the DP(α, G_0) prior is an asset — it allows ties in the θ_i, and thus makes DP mixture models appealing for many applications, including density estimation, classification, and regression

- Using the constructive definition of the DP, $G = \sum_{\ell=1}^{\infty} \omega_\ell \delta_{\vartheta_\ell}$, the prior probability model $f(\cdot; G, \phi)$ admits an (almost sure) representation as a countable mixture of parametric densities,

$$f(\cdot; G, \phi) = \sum_{\ell=1}^{\infty} \omega_\ell k(\cdot; \vartheta_\ell, \phi)$$

  $\rightarrow$ *weights*: $\omega_1 = z_1$, $\omega_\ell = z_\ell \prod_{r=1}^{\ell-1}(1 - z_r)$, $\ell \geq 2$, with z_r i.i.d. Beta(1,α)
  $\rightarrow$ *locations*: ϑ_ℓ i.i.d. G_0
  (and the sequences $\{z_r,\ r = 1,2,...\}$ and $\{\vartheta_\ell,\ \ell = 1,2,...\}$ are independent)

- This formulation has motivated study of several variants of the DP mixture model

- This formulation also helps motivate a link between limits of finite mixtures, with prior for the weights given by a symmetric Dirichlet distribution, and DP mixture models (e.g., Ishwaran & Zarepour, 2000)
- Consider the K finite mixture model

$$\sum_{t=1}^{K} \omega_t k(y; \vartheta_t)$$

  with $(\omega_1, \ldots, \omega_K) \sim \text{Dir}(\alpha/K, \ldots, \alpha/K)$ and $\vartheta_t \overset{i.i.d.}{\sim} G_0$, $t = 1, ..., K$
- When $K \to \infty$ this model can be shown to correspond to a DP mixture model with kernel k and a $\text{DP}(\alpha, G_0)$ prior for the mixing distribution

2.3 Prior specification

- Taking expectation over G with respect to its DP prior $\mathrm{DP}(\alpha, G_0)$, $\mathrm{E}\{F(\cdot; G, \phi)\} = F(\cdot; G_0, \phi)$, and $\mathrm{E}\{f(\cdot; G, \phi)\} = f(\cdot; G_0, \phi)$
- These expressions facilitate prior specification for the parameters ψ of $G_0(\cdot \mid \psi)$
- Recall that for the $\mathrm{DP}(\alpha, G_0)$ prior, α controls how *close* a realization G is to G_0
- In the DP mixture model, α controls the distribution of the number of distinct elements n^* of the vector $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$, and hence the number of distinct components of the mixture (Antoniak, 1974; Escobar & West, 1995; Liu, 1996)

Prior specification

- In particular,

$$\Pr(n^* = m \mid \alpha) = c_n(m) n! \alpha^m \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)}, \quad m = 1, ..., n,$$

  where the factors $c_n(m) = \Pr(n^* = m \mid \alpha = 1)$ can be computed using certain recurrence formulas (Stirling numbers) (Escobar & West, 1995)

- If α is assigned a prior $p(\alpha)$, $\Pr(n^* = m) = \int \Pr(n^* = m \mid \alpha) p(\alpha) \mathrm{d}\alpha$

- Moreover, for moderately large n,

$$\mathrm{E}(n^* \mid \alpha) \approx \alpha \log\left(\frac{\alpha + n}{\alpha}\right)$$

  and

$$\mathrm{Var}(n^* \mid \alpha) \approx \alpha \left\{\log\left(\frac{\alpha + n}{\alpha}\right) - 1\right\}$$

  which can be further averaged over the prior for α to obtain, for instance, a prior estimate for $\mathrm{E}(n^*)$

Prior specification

- Two *limiting* special cases of the DP mixture model

- One distinct component, when $\alpha \to 0^+$

$$\begin{aligned} y_i \mid \theta, \phi &\overset{ind.}{\sim} k(y_i; \theta, \phi), \quad i = 1, ..., n \\ \theta \mid \psi &\sim G_0(\cdot \mid \psi) \\ \phi, \psi &\sim p(\phi)p(\psi) \end{aligned}$$

- n components (one associated with each observation), when $\alpha \to \infty$

$$\begin{aligned} y_i \mid \theta_i, \phi &\overset{ind.}{\sim} k(y_i; \theta_i, \phi), \quad i = 1, ..., n \\ \theta_i \mid \psi &\overset{i.i.d.}{\sim} G_0(\cdot \mid \psi), \quad i = 1, ..., n \\ \phi, \psi &\sim p(\phi)p(\psi) \end{aligned}$$

2.4 Methods for posterior inference

- Data $= \{y_i, i = 1, ..., n\}$, i.i.d., conditionally on G and ϕ, from $f(\cdot; G, \phi)$ (if the model includes a regression component, the data also include the covariate vectors $\boldsymbol{x}_i$, and, in such cases, ϕ, typically, includes the vector of regression coefficients)
- Interest in inference for the latent mixing parameters $\boldsymbol{\theta} = (\theta_1, ..., \theta_n)$, for ϕ (and the hyperparameters α, ψ), for $f(y_0; G, \phi)$, and, in general, for functionals $H(F(\cdot; G, \phi))$ of the random mixture $F(\cdot; G, \phi)$ (e.g., cdf function, hazard function, mean and variance functionals, percentile functionals)
- Full and exact inference, given the data, for all these random quantities is based on the joint posterior of the DP mixture model

$$p(G, \phi, \boldsymbol{\theta}, \alpha, \psi \mid \text{data})$$

2.4.1 Marginal posterior simulation methods

- Key result: representation of the joint posterior (Antoniak, 1974)

$$p(G, \phi, \boldsymbol{\theta}, \alpha, \psi \mid \text{data}) = p(G \mid \boldsymbol{\theta}, \alpha, \psi) p(\boldsymbol{\theta}, \phi, \alpha, \psi \mid \text{data})$$

$\rightarrow p(\boldsymbol{\theta}, \phi, \alpha, \psi \mid \text{data})$ is the marginal posterior for the finite-dimensional portion of the full *parameter vector* $(G, \phi, \boldsymbol{\theta}, \alpha, \psi)$

$\rightarrow G \mid \boldsymbol{\theta}, \alpha, \psi \sim \text{DP}(\tilde{\alpha}, \tilde{G}_0)$, where $\tilde{\alpha} = \alpha + n$, and

$$\tilde{G}_0(\cdot) = \frac{\alpha}{\alpha + n} G_0(\cdot \mid \psi) + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{\theta_i}(\cdot)$$

(hence, the cdf, $\tilde{G}_0(t) = \frac{\alpha}{\alpha+n} G_0(t \mid \psi) + \frac{1}{\alpha+n} \sum_{i=1}^{n} 1_{[\theta_i, \infty)}(t)$)

- Sampling from the $\text{DP}(\tilde{\alpha}, \tilde{G}_0)$ is possible using one of its definitions — thus, we can obtain full posterior inference under DP mixture models if we can sample from the marginal posterior $p(\boldsymbol{\theta}, \phi, \alpha, \psi \mid \text{data})$

Methods for posterior inference

- The marginal posterior $p(\boldsymbol{\theta}, \phi, \alpha, \psi \mid \text{data})$ corresponds to the marginalized version of the DP mixture model, obtained after integrating G over its DP prior (Blackwell & MacQueen, 1973),

$$\begin{aligned} y_i \mid \theta_i, \phi &\overset{ind.}{\sim} k(y_i; \theta_i, \phi), \quad i = 1, ..., n \\ \boldsymbol{\theta} = (\theta_1, ..., \theta_n) \mid \alpha, \psi &\sim p(\boldsymbol{\theta} \mid \alpha, \psi) \\ \phi, \alpha, \psi &\sim p(\phi)p(\alpha)p(\psi) \end{aligned}$$

- The induced prior distribution $p(\boldsymbol{\theta} \mid \alpha, \psi)$ for the mixing parameters θ_i can be developed by exploiting the Pólya urn characterization of the DP,

$$p(\boldsymbol{\theta} \mid \alpha, \psi) = G_0(\theta_1 \mid \psi) \prod_{i=2}^{n} \left\{ \frac{\alpha}{\alpha + i - 1} G_0(\theta_i \mid \psi) + \frac{1}{\alpha + i - 1} \sum_{j=1}^{i-1} \delta_{\theta_j}(\theta_i) \right\}$$

  $\rightarrow$ for increasing sample sizes, the joint prior $p(\boldsymbol{\theta} \mid \alpha, \psi)$ gets increasingly complex to work with

Methods for posterior inference

- Therefore, the marginal posterior

$$p(\boldsymbol{\theta}, \phi, \alpha, \psi \mid \text{data}) \propto p(\boldsymbol{\theta} \mid \alpha, \psi) p(\phi) p(\alpha) p(\psi) \prod_{i=1}^{n} k(y_i; \theta_i, \phi)$$

  is difficult to work with — even point estimates practically impossible to compute for moderate to large sample sizes

- Early work for posterior inference:
  → some results for certain problems in density estimation, i.e., expressions for Bayes point estimates of $f(y_0; G)$ (Lo, 1984; Brunner & Lo, 1989)
  → approximations for special cases, e.g., for binomial DP mixtures (Berry & Christensen, 1979)
  → Monte Carlo integration algorithms to obtain point estimates for the θ_i (Ferguson, 1983; Kuo, 1986a,b)

Simulation-based model fitting

- Note that, although the joint prior $p(\boldsymbol{\theta} \mid \alpha, \psi)$ has an awkward expression for samples of realistic size n, the prior full conditionals have convenient expressions

$$p(\theta_i \mid \{\theta_j : j \neq i\}, \alpha, \psi) = \frac{\alpha}{\alpha + n - 1} G_0(\theta_i \mid \psi) + \frac{1}{\alpha + n - 1} \sum_{j=1}^{n-1} \delta_{\theta_j}(\theta_i)$$

- **Key idea**: (Escobar, 1988; 1994) setup a Markov chain to explore the posterior $p(\boldsymbol{\theta}, \phi, \alpha, \psi \mid \text{data})$ by simulating only from posterior full conditional distributions, which arise by combining the likelihood terms with the corresponding prior full conditionals
  (in fact, Escobar's algorithm is essentially a Gibbs sampler developed for a specific class of models!)

- Several other Markov chain Monte Carlo (MCMC) methods that improve on the original algorithm (e.g., West et al., 1994; Escobar & West, 1995; Bush & MacEachern, 1996; Neal, 2000; Jain & Neal, 2004)

Methods for posterior inference

- A key property for the implementation of the Gibbs sampler is the discreteness of G, which induces a clustering of the θ_i
  $\rightarrow$ n^*: number of distinct elements (clusters) in the vector $(\theta_1, ..., \theta_n)$
  $\rightarrow$ θ_j^*, $j = 1,...,n^*$: the distinct θ_i
  $\rightarrow$ $\boldsymbol{w} = (w_1, ..., w_n)$: vector of configuration indicators, defined by $w_i = j$ if and only if $\theta_i = \theta_j^*$, $i = 1,...,n$
  $\rightarrow$ n_j: size of j-th cluster, i.e., $n_j = |\{i : w_i = j\}|$, $j = 1,...,n^*$

- Evidently, $(n^*, \boldsymbol{w}, (\theta_1^*, ..., \theta_{n^*}^*))$ yields an equivalent representation for $(\theta_1, ..., \theta_n)$

- Standard Gibbs sampler to draw from $p(\boldsymbol{\theta}, \phi, \alpha, \psi \mid \text{data})$ (Escobar, 1994; Escobar & West, 1995) is based on the following full conditionals:
  (a) $p(\theta_i \mid \{\theta_{i'} : i' \neq i\}, \alpha, \psi, \phi, \text{data})$, for $i = 1,...,n$
  (b) $p(\alpha \mid n^*, \text{data})$ and $p(\psi \mid \{\theta_j^*, j = 1, ..., n^*\}, n^*)$
  (c) $p(\phi \mid \{\theta_i : i = 1, ..., n\}, \text{data})$
  (the expressions include conditioning only on the relevant variables, exploiting the conditional independence structure of the model and properties of the DP)

- (a) For each $i = 1, ..., n$

$$p(\theta_i \mid \{\theta_{i'} : i' \neq i\}, \alpha, \psi, \phi, \text{data}) = \frac{q_0 h(\theta_i \mid \psi, \phi, y_i) + \sum_{j=1}^{n^{*-}} n_j^- q_j \delta_{\theta_j^{*-}}(\theta_i)}{q_0 + \sum_{j=1}^{n^{*-}} n_j^- q_j}$$

$\rightarrow q_j = k(y_i; \theta_j^{*-}, \phi)$
$\rightarrow q_0 = \alpha \int k(y_i; \theta, \phi) g_0(\theta \mid \psi) \mathrm{d}\theta$
$\rightarrow h(\theta_i \mid \psi, \phi, y_i) \propto k(y_i; \theta_i, \phi) g_0(\theta_i \mid \psi)$
$\rightarrow g_0$ is the density of G_0
$\rightarrow$ superscript "$-$" denotes all relevant quantities when θ_i is removed from the vector $(\theta_1, ..., \theta_n)$, e.g., n^{*-} is the number of clusters in $\{\theta_{i'} : i' \neq i\}$

- Note that updating θ_i implicitly updates w_i, $i = 1,...,n$ — before updating θ_{i+1}, we redefine n^*, θ_j^*, $j = 1,...,n^*$, w_i, $i = 1,...,n$, and n_j, $j = 1,...,n^*$

- (b) Although the posterior full conditional for α is not of a standard form, an augmentation method facilitates sampling provided the prior for α is a gamma distribution (say, with mean a_α/b_α) (Escobar & West, 1995),

$$\begin{aligned} p(\alpha \mid n^*, \text{data}) &\propto p(\alpha)\alpha^{n^*}\tfrac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \\ &\propto p(\alpha)\alpha^{n^*-1}(\alpha+n)\text{Beta}(\alpha+1,n) \\ &\propto p(\alpha)\alpha^{n^*-1}(\alpha+n)\textstyle\int_0^1 x^\alpha(1-x)^{n-1}dx \end{aligned}$$

$\rightarrow$ introduce an auxiliary variable η such that

$$p(\alpha,\eta \mid n^*, \text{data}) \propto p(\alpha)\alpha^{n^*-1}(\alpha+n)\eta^\alpha(1-\eta)^{n-1},$$

$\rightarrow$ extend the Gibbs sampler to draw from $p(\eta \mid \alpha, \text{data}) = \text{Beta}(\alpha+1, n)$, and $p(\alpha \mid \eta, n^*, \text{data})$, which is given by the two-component mixture

$$p\text{gamma}(a_\alpha + n^*, b_\alpha - \log(\eta)) + (1-p)\text{gamma}(a_\alpha + n^* - 1, b_\alpha - \log(\eta))$$

where $p = (a_\alpha + n^* - 1)/\{n(b_\alpha - \log(\eta)) + a_\alpha + n^* - 1\}$

- Regarding the parameters ψ of G_0,

$$p(\psi \mid \{\theta_j^*, j = 1, ..., n^*\}, n^*) \propto p(\psi) \prod_{j=1}^{n^*} g_0(\theta_j^* \mid \psi)$$

  leading, typically, to standard updates

- (c) The posterior full conditional for ϕ does not involve the nonparametric part of the DP mixture model,

$$p(\phi \mid \{\theta_i : i = 1, ..., n\}, \text{data}) \propto p(\phi) \prod_{i=1}^{n} k(y_i; \theta_i, \phi)$$

- **Improved Gibbs sampler** (West et al., 1994; Bush & MacEachern, 1996): adds one more step where the cluster locations θ_j^* are resampled at each iteration to improve the mixing of the chain
  $\rightarrow$ at each iteration, once step (a) is completed, we obtain a specific number of clusters n^*, and a specific configuration $\boldsymbol{w} = (w_1, ..., w_n)$
  $\rightarrow$ after the marginalization over G, the prior for the θ_j^*, given the partition $(n^*, \boldsymbol{w})$, is given by

$$p(\theta_j^* : j = 1, ..., n^* \mid n^*, \boldsymbol{w}, \psi) = \prod_{j=1}^{n^*} g_0(\theta_j^* \mid \psi)$$

  i.e., given n^* and $\boldsymbol{w}$, the θ_j^* are i.i.d. from G_0

  $\rightarrow$ hence, for each $j = 1,...,n^*$, the posterior full conditional

$$p(\theta_j^* \mid \boldsymbol{w}, n^*, \psi, \phi, \text{data}) \propto g_0(\theta_j^* \mid \psi) \prod_{\{i:w_i=j\}} k(y_i; \theta_j^*, \phi)$$

Methods for posterior inference

- **Note**: the Gibbs sampler can be difficult or inefficient to implement if
  $\rightarrow$ the integral $\int k(y;\theta,\phi)g_0(\theta \mid \psi)\mathrm{d}\theta$ is not available in closed form (and numerical integration is not feasible or reliable)
  and/or
  $\rightarrow$ random generation from $h(\theta \mid \psi,\phi,y) \propto k(y;\theta,\phi)g_0(\theta \mid \psi)$ is not readily available

- For such cases, alternative MCMC algorithms have been proposed in the literature (e.g., MacEachern & Müller, 1998; Neal, 2000; Dahl, 2005; Jain & Neal, 2007)

- Extensions for data structures that include missing or censored observations are also possible (Kuo & Smith, 1992; Kuo & Mallick, 1997; Kottas, 2006b)

- Alternative (to MCMC) fitting techniques have also been studied (e.g., Liu, 1996; MacEachern et al., 1999; Newton & Zhang, 1999; Blei & Jordan, 2006)

Posterior predictive distributions

- Implementing one of the available MCMC algorithms for DP mixture models, we obtain B posterior samples

$$\{\boldsymbol{\theta}_b = (\theta_{ib} : i = 1, ..., n), \alpha_b, \psi_b, \phi_b\}, \quad b = 1, ..., B,$$

from $p(\boldsymbol{\theta}, \phi, \alpha, \psi \mid \text{data})$, equivalently, posterior samples

$$\left\{n_b^*, \boldsymbol{w}_b, \boldsymbol{\theta}_b^* = (\theta_{jb}^* : j = 1, ..., n_b^*), \alpha_b, \psi_b, \phi_b\right\}, \quad b = 1, ..., B,$$

from $p(n^*, \boldsymbol{w}, \boldsymbol{\theta}^* = (\theta_j^* : j = 1, ..., n^*), \phi, \alpha, \psi \mid \text{data})$

- Bayesian *density estimate* is based on the posterior predictive density $p(y_0 \mid \text{data})$ corresponding to a *new* y_0 with associated mixing parameter θ_0
- Using, again, the Pólya urn structure for the DP,

$$p(\theta_0 \mid n^*, \boldsymbol{w}, \boldsymbol{\theta}^*, \alpha, \psi) = \frac{\alpha}{\alpha + n} G_0(\theta_0 \mid \psi) + \frac{1}{\alpha + n} \sum_{j=1}^{n^*} n_j \delta_{\theta_j^*}(\theta_0)$$

Methods for posterior inference

- The posterior predictive distribution for y_0 is given by

$$\begin{aligned} p(y_0 \mid \text{data}) &= \int p(y_0 \mid n^*, \boldsymbol{w}, \boldsymbol{\theta}^*, \alpha, \psi, \phi) p(n^*, \boldsymbol{w}, \boldsymbol{\theta}^*, \alpha, \psi, \phi \mid \text{data}) \\ &= \int \int p(y_0 \mid \theta_0, \phi) p(\theta_0 \mid n^*, \boldsymbol{w}, \boldsymbol{\theta}^*, \alpha, \psi) p(n^*, \boldsymbol{w}, \boldsymbol{\theta}^*, \alpha, \psi, \phi \mid \text{data}) \end{aligned}$$

  $\rightarrow$ hence, a sample $(y_{0,b} : b = 1, ..., B)$ from $p(y_0 \mid \text{data})$ can be obtained using the MCMC output: for each $b = 1, ..., B$, we first draw $\theta_{0,b}$ from $p(\theta_0 \mid n_b^*, \boldsymbol{w}_b, \boldsymbol{\theta}_b^*, \alpha_b, \psi_b)$, and then draw $y_{0,b}$ from $p(y_0 \mid \theta_{0,b}, \phi_b) = K(\cdot; \theta_{0,b}, \phi_b)$

- To further highlight the mixture structure, note that we can also write

$$p(y_0 \mid \text{data}) = \int \left(\frac{\alpha}{\alpha+n} \int k(y_0 \mid \theta, \phi) g_0(\theta \mid \psi) \mathrm{d}\theta + \frac{1}{\alpha+n} \sum_{j=1}^{n^*} n_j k(y_0; \theta_j^*, \phi) \right) p(n^*, \boldsymbol{w}, \boldsymbol{\theta}^*, \alpha, \psi, \phi \mid \text{data})$$

  $\rightarrow$ the integrand above is a mixture with $n^* + 1$ components — the last n^* components (that dominate when α is small relative to n) yield a discrete mixture (in θ) of $k(\cdot; \theta, \phi)$ with the mixture parameters defined by the distinct θ_j^* — the posterior predictive density for y_0 is obtained by averaging this mixture with respect to the posterior of n^*, $\boldsymbol{w}$, $\boldsymbol{\theta}^*$ and all other parameters

Inference for general functionals of the random mixture

- Note that $p(y_0 \mid \text{data})$ is the posterior point estimate for the density functional $f(y_0; G, \phi)$ (at point y_0), i.e., $p(y_0 \mid \text{data}) = \text{E}(f(y_0; G, \phi) \mid \text{data})$ (the Bayesian density estimate under a DP mixture model can be obtained without sampling from the posterior of G)

- Analogously, we can obtain posterior moments for linear functionals $H(F(\cdot; G, \phi)) = \int H(K(\cdot; \theta, \phi))\text{d}G(\theta)$ (Gelfand & Mukhopadhyay, 1995) — for linear functionals, the functional of the mixture is the mixture of the functionals applied to the parametric kernel (e.g., density and cdf functionals, mean functional)

- How about more general inference for functionals?
  $\rightarrow$ interval estimates for $F(y_0; G, \phi)$ for specified y_0, and, therefore, (pointwise) uncertainty bands for $F(\cdot; G, \phi)$?
  $\rightarrow$ inference for derived functions from $F(\cdot; G, \phi)$, e.g., cumulative hazard, $-\log(1 - F(\cdot; G, \phi))$, or hazard, $f(\cdot; G, \phi)/(1 - F(\cdot; G, \phi))$, functions?
  $\rightarrow$ inference for non-linear functionals, e.g., for median, and general percentiles?

Methods for posterior inference

- Such inferences require the posterior of G — recall,

$$p(G, \phi, \boldsymbol{\theta}, \alpha, \psi \mid \text{data}) = p(G \mid \boldsymbol{\theta}, \alpha, \psi) p(\boldsymbol{\theta}, \phi, \alpha, \psi \mid \text{data})$$

$$\text{and } G \mid \boldsymbol{\theta}, \alpha, \psi \sim \text{DP}(\alpha + n, \tilde{G}_0 = \alpha(\alpha + n)^{-1} G_0(\psi) + (\alpha + n)^{-1} \textstyle\sum_{i=1}^{n} \delta_{\theta_i}),$$

- Hence, after fitting the marginalized version of the DP mixture model (i.e., with G integrated out over its DP prior), using one of the available MCMC methods, we can obtain draws from the posterior of G

- For each posterior sample $(\boldsymbol{\theta}_b, \alpha_b, \psi_b, \phi_b)$, $b = 1, ..., B$, we can draw G_b from $p(G \mid \boldsymbol{\theta}_b, \alpha_b, \psi_b)$ using:
  $\rightarrow$ the constructive definition of the DP with a truncation approximation (Gelfand & Kottas, 2002; Kottas, 2006b)
  $\rightarrow$ the original DP definition if we only need sample paths for the cdf of the mixture (and y is univariate) (e.g., Krnjajić, Kottas & Draper, 2008)

- Finally, the posterior samples G_b yield posterior samples $\{H(F(\cdot; G_b, \phi_b)) : b = 1, ..., B\}$ from any functional $H(F(\cdot; G, \phi))$ of interest

Density estimation data example

- As an example, we analyze the `galaxy` data set: velocities (km/second) for 82 galaxies, drawn from six well-separated conic sections of the Corona Borealis region
- The model is a location-scale DP mixture of Gaussian distributions, with a conjugate normal-inverse gamma baseline distribution:

$$f(\cdot; G) = \int \mathrm{N}(\cdot; \mu, \sigma^2) \mathrm{d}G(\mu, \sigma^2), \quad G \sim \mathrm{DP}(\alpha, G_0)$$

  where $G_0(\mu, \sigma^2) = \mathrm{N}(\mu; \mu_0, \sigma^2/\kappa)\mathrm{IGamma}(\sigma^2; \nu, s)$
- We consider four different prior specifications to explore the effect of increasing flexibility in the DP prior hyperparameters
- Figure 2.1 shows posterior predictive density estimates obtained using the function `DPdensity` in the `R` package `DPpackage` (the code employed is one of the examples in the help file)

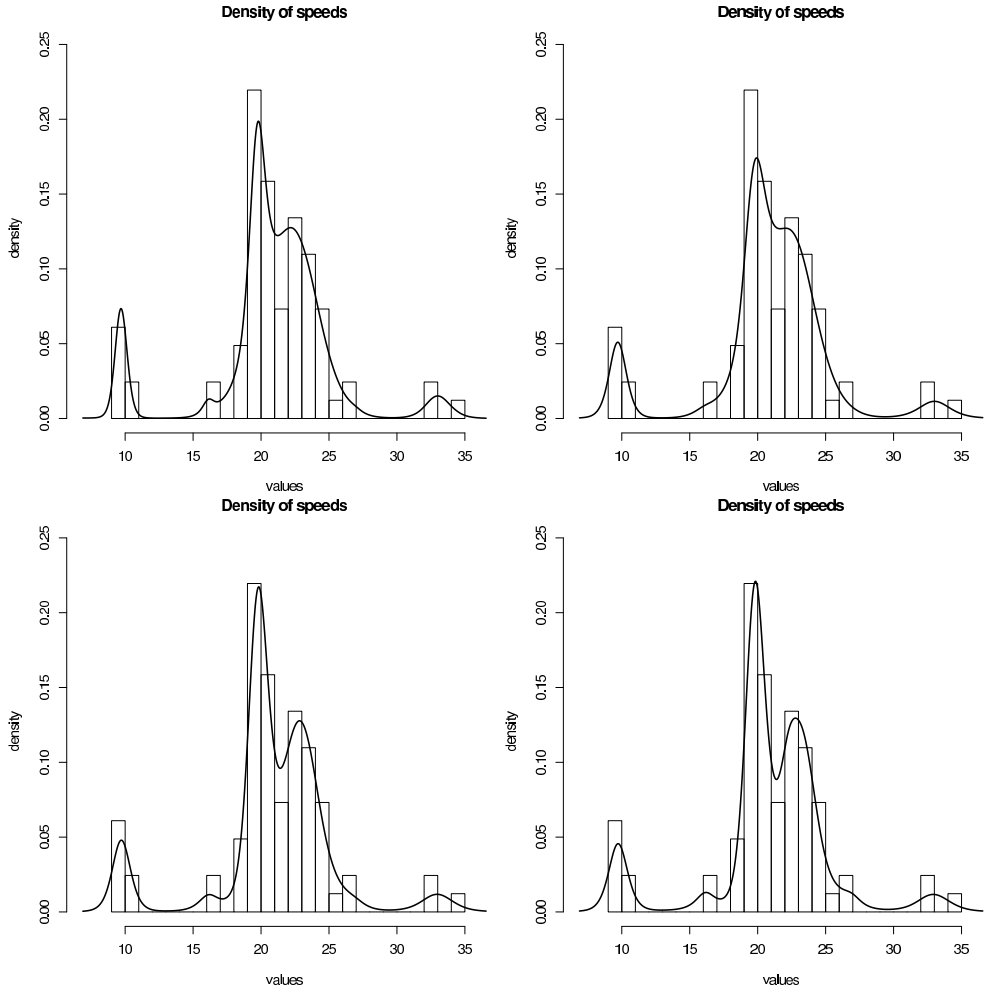


Figure 2.1: Histograms of the raw data and posterior predictive densities under four prior choices for the `galaxy` data. In the top left panel we set $\alpha = 1$, $\mu_0 = 0$, $s = 2$, $\nu = 4$, $\kappa \sim \text{gamma}(0.5, 50)$; the top right panel uses the same settings except $s \sim \text{IGamma}(4, 2)$; in the bottom left panel we add hyperprior $\mu_0 \sim \text{N}(0, 100000)$; and in the bottom right panel we further add hyperprior $\alpha \sim \text{gamma}(2, 2)$.

2.4.2 Conditional posterior simulation methods

- The main characteristic of the MCMC methods of Section 4.1 is that they are based on the marginal posterior of the DP mixture model, $p(\boldsymbol{\theta}, \phi, \alpha, \psi \mid \text{data})$, resulting after marginalizing the random mixing distribution G (thus, referred to as *marginal* methods)
- Although posterior inference for G is still possible, it is of interest to study alternative *conditional* posterior simulation approaches that impute G as a component of the MCMC algorithm
- Most of the emphasis on conditional methods based on finite truncation approximation of G, using its stick-breaking representation — main example: Blocked Gibbs sampler (Ishwaran & Zarepour, 2000; Ishwaran & James, 2001)
- More recent work based on retrospective sampling techniques (Papaspiliopoulos & Roberts, 2008)

Methods for posterior inference

- **Blocked Gibbs sampling:** based on truncation approximation to mixing distribution G given, for finite N, by

$$G_N(\cdot) = \sum_{\ell=1}^{N} p_\ell \delta_{Z_\ell}(\cdot)$$

$\rightarrow$ the Z_ℓ, $\ell = 1, ..., N$, are i.i.d. G_0

$\rightarrow$ the weights arise through stick-breaking (with truncation)

$$p_1 = V_1; \;\; p_\ell = V_\ell \prod_{r=1}^{\ell-1}(1 - V_r), \; \ell = 2, ..., N-1; \;\; p_N = \prod_{r=1}^{N-1}(1 - V_r)$$

where the V_ℓ, $\ell = 1, ..., N-1$, are i.i.d. Beta$(1, \alpha)$

$\rightarrow$ choice of N follows guidelines discussed earlier

- The joint prior for $\boldsymbol{p} = (p_1, ..., p_N)$, given α, corresponds to a special case of the generalized Dirichlet distribution (Connor & Mosimann, 1969),

$$f(\boldsymbol{p} \mid \alpha) = \alpha^{N-1} p_N^{\alpha-1} (1 - p_1)^{-1} (1 - (p_1 + p_2))^{-1} \times ... \times (1 - \sum_{\ell=1}^{N-2} p_\ell)^{-1}$$

- Replacing G with $G_N \equiv (\boldsymbol{p}, \boldsymbol{Z})$, where $\boldsymbol{Z} = (Z_1, ..., Z_N)$, in the generic DP mixture model hierarchical formulation, we have:

$$\begin{aligned} y_i \mid \theta_i, \phi &\overset{ind.}{\sim} k(y_i; \theta_i, \phi), \quad i = 1, ..., n \\ \theta_i \mid \boldsymbol{p}, \boldsymbol{Z} &\overset{i.i.d.}{\sim} G_N, \quad i = 1, ..., n \\ \boldsymbol{p}, \boldsymbol{Z} \mid \alpha, \psi &\sim f(\boldsymbol{p} \mid \alpha) \prod_{\ell=1}^{N} g_0(Z_\ell \mid \psi) \\ \phi, \alpha, \psi &\sim p(\phi)p(\alpha)p(\psi) \end{aligned}$$

- If we marginalize over the θ_i in the first two stages of the hierarchical model, we obtain a finite mixture model for the y_i,

$$f(\cdot; \boldsymbol{p}, \boldsymbol{Z}, \phi) = \sum_{\ell=1}^{N} p_\ell k(\cdot; Z_\ell, \phi)$$

(conditionally on $(\boldsymbol{p}, \boldsymbol{Z})$ and ϕ), which replaces the countable DP mixture, $f(\cdot; G, \phi) = \int k(\cdot; \theta, \phi) \mathrm{d}G(\theta) = \sum_{\ell=1}^{\infty} \omega_\ell k(\cdot; \vartheta_\ell, \phi)$

Methods for posterior inference

- Now, having approximated the countable DP mixture with a finite mixture, the mixing parameters θ_i can be replaced with configuration variables $\boldsymbol{L} = (L_1, ..., L_n)$ — each L_i takes values in $\{1, ..., N\}$ such that $L_i = \ell$ if-f $\theta_i = Z_\ell$, for $i = 1, ..., n$; $\ell = 1, ..., N$

- Final version of the hierarchical model:

$$\begin{aligned}
y_i \mid \boldsymbol{Z}, L_i, \phi &\overset{ind.}{\sim} k(y_i; Z_{L_i}, \phi), \quad i = 1, ..., n \\
L_i \mid \boldsymbol{p} &\overset{i.i.d.}{\sim} \textstyle\sum_{\ell=1}^{N} p_\ell \delta_\ell(L_i), \quad i = 1, ..., n \\
\boldsymbol{p} \mid \alpha &\sim f(\boldsymbol{p} \mid \alpha) \\
Z_\ell \mid \psi &\overset{i.i.d.}{\sim} G_0(\cdot \mid \psi), \quad \ell = 1, ..., N \\
\phi, \alpha, \psi &\sim p(\phi)p(\alpha)p(\psi)
\end{aligned}$$

- Marginalizing over the L_i in the first two stages of the model, we obtain the same finite mixture model for the y_i: $f(\cdot; \boldsymbol{p}, \boldsymbol{Z}, \phi) = \sum_{\ell=1}^{N} p_\ell k(\cdot; Z_\ell, \phi)$

Simulation-based model fitting

- **Gibbs sampling** for posterior distribution $p(\boldsymbol{Z}, \boldsymbol{p}, \boldsymbol{L}, \phi, \alpha, \psi \mid \text{data})$

- Updating the Z_ℓ, $\ell = 1, ..., N$:
  $\rightarrow$ let n^* be the number of distinct values $\{L_j^* : j = 1, ..., n^*\}$ of vector $\boldsymbol{L}$
  $\rightarrow$ then, the posterior full conditional for Z_ℓ, $\ell = 1, ..., N$, can be expressed in general as:

$$p(Z_\ell \mid ..., \text{data}) \propto g_0(Z_\ell \mid \psi) \prod_{j=1}^{n^*} \prod_{\{i:L_i=L_j^*\}} k(y_i; Z_{L_j^*}, \phi)$$

  $\rightarrow$ if $\ell \notin \{L_j^* : j = 1, ..., n^*\}$, Z_ℓ is drawn from $G_0(\cdot \mid \psi)$
  $\rightarrow$ for $\ell = L_j^*$, $j = 1, ..., n^*$,

$$p(Z_{L_j^*} \mid ..., \text{data}) \propto g_0(Z_{L_j^*} \mid \psi) \prod_{\{i:L_i=L_j^*\}} k(y_i; Z_{L_j^*}, \phi)$$

Methods for posterior inference

- The posterior full conditional for $\boldsymbol{p}$: $p(\boldsymbol{p} \mid ..., \text{data}) \propto f(\boldsymbol{p} \mid \alpha) \prod_{\ell=1}^{N} p_\ell^{M_\ell}$, where $M_\ell = |\{i : L_i = \ell\}|$, $\ell = 1, ..., N$
  $\rightarrow$ results in a generalized Dirichlet distribution, which can be sampled through independent latent Beta variables
  $\rightarrow V_\ell^* \overset{ind.}{\sim} \text{Beta}(1 + M_\ell, \alpha + \sum_{r=\ell+1}^{N} M_r)$, $\ell = 1, ..., N-1$
  $\rightarrow p_1 = V_1^*$; $p_\ell = V_\ell^* \prod_{r=1}^{\ell-1}(1 - V_r^*)$, $\ell = 2, ..., N-1$; $p_N = 1 - \sum_{\ell=1}^{N-1} p_\ell$

- Updating the L_i, $i = 1, ..., n$:
  $\rightarrow$ each L_i is drawn from the discrete distribution on $\{1, ..., N\}$ with probabilities $\tilde{p}_{\ell i} \propto p_\ell k(y_i; Z_\ell, \phi)$, $\ell = 1, ..., N$

- **Note:** the update for each L_i does not depend on the other $L_{i'}$, $i' \neq i$ — this aspect of this Gibbs sampler, along with the *block updates* for the Z_ℓ, are key advantages over Pólya urn based marginal MCMC methods

Methods for posterior inference

- The posterior full conditionals for ϕ, α, and ψ also have convenient forms:

  $\rightarrow p(\phi \mid ..., \text{data}) \propto p(\phi) \prod_{i=1}^{n} k(y_i; \theta_i, \phi)$

  $\rightarrow p(\psi \mid ..., \text{data}) \propto p(\psi) \prod_{\ell=1}^{N} g_0(Z_\ell \mid \psi)$ (which can also be expressed in terms of only the distinct $Z_{L_j^*}$: $p(\psi \mid ..., \text{data}) \propto p(\psi) \prod_{j=1}^{n^*} g_0(Z_{L_j^*} \mid \psi)$)

  $\rightarrow p(\alpha \mid ..., \text{data}) \propto p(\alpha) \alpha^{N-1} p_N^{\alpha}$, which with a gamma$(a_\alpha, b_\alpha)$ prior for α, results in a gamma$(N + a_\alpha - 1, b_\alpha - \log p_N)$ full conditional (for numerical stability, compute $\log p_N = \log \prod_{r=1}^{N-1} (1 - V_r^*) = \sum_{r=1}^{N-1} \log(1 - V_r^*)$)

- The posterior samples from $p(\boldsymbol{Z}, \boldsymbol{p}, \boldsymbol{L}, \phi, \alpha, \psi \mid \text{data})$ yield directly the posterior for G_N, and thus, full posterior inference for any functional of the (approximate) DP mixture $f(\cdot; G_N, \phi) \equiv f(\cdot; \boldsymbol{p}, \boldsymbol{Z}, \phi)$

Posterior predictive inference

- Posterior predictive density for *new* y_0, with corresponding configuration variable L_0,

$$\begin{aligned} p(y_0 \mid \text{data}) &= \iint k(y_0; Z_{L_0}, \phi) \left(\sum_{\ell=1}^{N} p_\ell \delta_\ell(L_0) \right) p(\boldsymbol{Z}, \boldsymbol{p}, \boldsymbol{L}, \phi, \alpha, \psi \mid \text{data}) \\ &\qquad \mathrm{d}L_0 \, \mathrm{d}\boldsymbol{Z} \, \mathrm{d}\boldsymbol{L} \, \mathrm{d}\boldsymbol{p} \, \mathrm{d}\phi \, \mathrm{d}\alpha \, \mathrm{d}\psi \\ &= \int \left(\sum_{\ell=1}^{N} p_\ell k(y_0; Z_\ell, \phi) \right) p(\boldsymbol{Z}, \boldsymbol{p}, \boldsymbol{L}, \phi, \alpha, \psi \mid \text{data}) \\ &\qquad \mathrm{d}\boldsymbol{Z} \, \mathrm{d}\boldsymbol{L} \, \mathrm{d}\boldsymbol{p} \, \mathrm{d}\phi \, \mathrm{d}\alpha \, \mathrm{d}\psi \\ &= \mathrm{E}(f(y_0; \boldsymbol{p}, \boldsymbol{Z}, \phi) \mid \text{data}) \end{aligned}$$

- Hence, $p(y_0 \mid \text{data})$ can be estimated over a grid in y_0 by drawing samples $\{L_{0b} : b = 1, ..., B\}$ for L_0, based on the posterior samples for $\boldsymbol{p}$, and computing the Monte Carlo estimate $B^{-1} \sum_{b=1}^{B} k(y_0; Z_{L_{0b}}, \phi_b)$, where B is the posterior sample size

Notes 3: Dirichlet process mixture models – Applications

Outline

3.1 Summary and references

3.2 Survival analysis using Weibull DP mixtures

3.3 Curve fitting using Dirichlet process mixtures

3.4 Modelling for multivariate ordinal data

3.5 Nonparametric inference for Poisson processes

3.6 Modelling for stochastically ordered distributions

3.1 Summary and references

Dirichlet process (DP) mixture models, and their extensions, have largely dominated methodological and applied Bayesian nonparametric work, after the technology for their simulation-based model fitting was introduced.
References categorized by methodological/application area include:

- Models for binary and ordinal data (Erkanli et al., 1993; Basu & Mukhopadhyay, 2000; Hoff, 2005; Das & Chattopadhyay, 2004; Kottas et al., 2005)
- Density estimation, mixture deconvolution, and curve fitting (West et al., 1994; Escobar & West, 1995; Cao & West, 1996; Gasparini, 1996; Müller et al., 1996; Ishwaran & James, 2002; Do et al., 2005; Leslie et al., 2007; Lijoi et al., 2007)
- Regression modelling with structured error distributions and/or regression functions (Brunner, 1995; Lavine & Mockus, 1995; Kottas & Gelfand, 2001b; Dunson, 2005; Kottas & Krnjajić, 2009)

Summary and references

- Regression models for survival/reliability data (Kuo & Mallick, 1997; Gelfand & Kottas, 2003; Merrick et al., 2003; Hanson, 2006b)
- Generalized linear, and linear mixed, models (Bush & MacEachern, 1996; Kleinman & Ibrahim, 1998; Mukhopadhyay & Gelfand, 1997; Müller & Rosner, 1997; Quintana, 1998)
- Errors-in-variables models (Müller & Roeder, 1997); Multiple comparisons problems (Gopalan & Berry, 1998); Analysis of selection models (Lee & Berger, 1999)
- Meta-analysis and nonparametric ANOVA models (Mallick & Walker, 1997; Tomlinson & Escobar, 1999; Burr et al., 2003; De Iorio et al., 2004; Müller et al., 2004; Müller et al., 2005)
- Time series modelling and econometrics applications (Müller et al., 1997; Chib & Hamilton, 2002; Hirano, 2002; Hasegawa & Kozumi, 2003; Griffin & Steel, 2004)
- ROC data analysis (Erkanli et al., 2006; Hanson et al., 2008)

3.2 Survival analysis using Weibull DP mixtures

- Bayesian nonparametric work for survival analysis has focused on prior models for cumulative hazard or hazard functions (gamma processes, extended gamma processes, Beta processes), or survival functions (Dirichlet processes, Polya trees) (see, e.g., Walker et al., 1999; Ibrahim et al., 2001)

- Dirichlet process mixture models?

$$f(t;G) = \int k(t;\theta)\mathrm{d}G(\theta),\ t \in R^+, \quad G \sim \mathrm{DP}(\alpha, G_0)$$

  $\rightarrow$ kernel $k(t;\theta)$ that yields mixtures with flexible density (and hazard) shapes is needed

- Mixtures of Weibull or gamma distributions (Kottas, 2006b; Hanson, 2006b)

- **Weibull Dirichlet process mixture model**

$$
\begin{aligned}
t_i \mid (\gamma_i, \lambda_i) &\overset{ind.}{\sim} K(t \mid \gamma_i, \lambda_i) = 1 - \exp(-t^{\gamma_i}/\lambda_i), \;\; i = 1, ..., n \\
(\gamma_i, \lambda_i) \mid G &\overset{i.i.d.}{\sim} G, \;\; i = 1, ..., n \\
G \mid \alpha, \phi, \psi &\sim \mathrm{DP}(\alpha, G_0); G_0 = U(\gamma \mid 0, \phi) IG(\lambda \mid c, \psi) \\
\alpha, \phi, \psi &\sim p(\alpha) p(\phi) p(\psi)
\end{aligned}
$$

- full posterior inference for all functionals of interest in survival analysis, including non-linear functionals (e.g., hazard function, and median survival time) — uses sampling from the posterior of G
- also, the prior distribution of functionals can be sampled (quantifies prior to posterior learning)

Data Illustrations

- Simulated data ($n = 200$) from a mixture of two Lognormal distributions $0.8LN(0, 0.25) + 0.2LN(1.2, 0.02)$
  $\rightarrow$ bimodal density
  $\rightarrow$ non-monotone hazard function with 3 change points in the interval $(0, 5)$ where essentially all the probability mass lies

- Remission times (in weeks) for leukemia patients (Lawless, 1982)
  $\rightarrow$ comparison of two treatments, A and B, each with 20 patients (3 and 2 right censored survival times, respectively)
  $\rightarrow$ "no evidence of a difference in distributions" based on classical tests that rely on approximate normality **and** assume proportional hazard functions (Lawless, 1982)

Survival analysis using Weibull DP mixtures

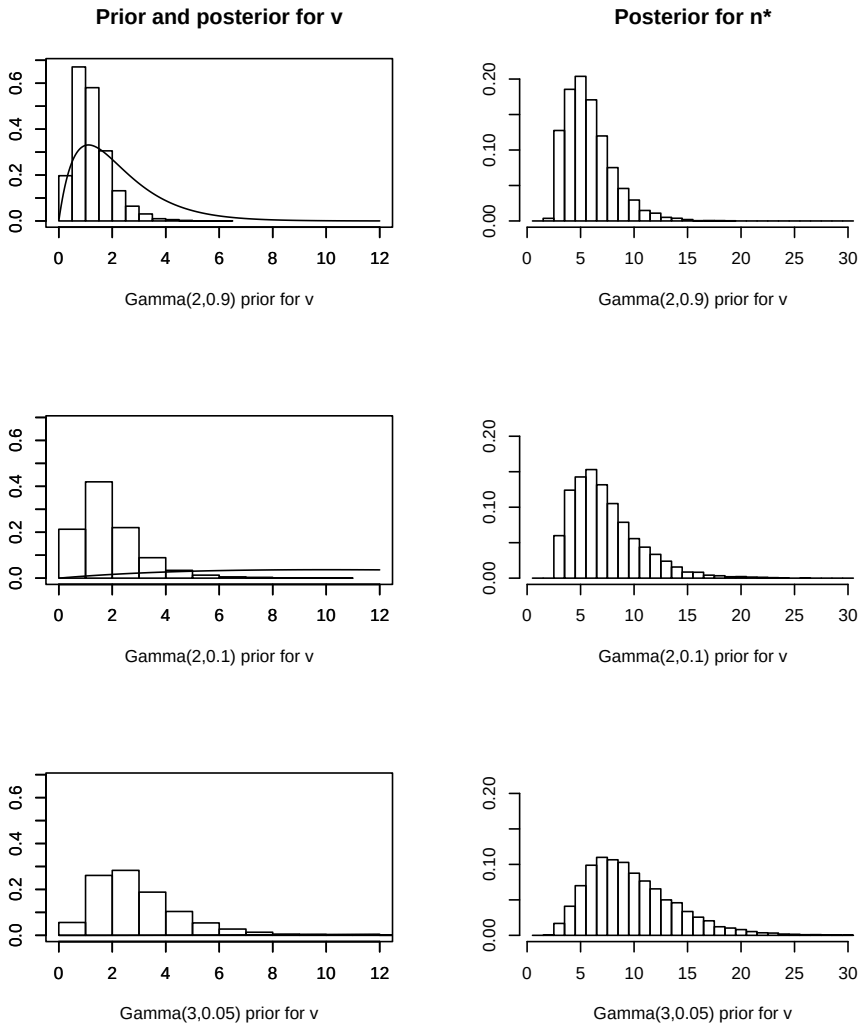


Figure 3.1: Simulated data. Histograms of posterior draws for α (denoted by v in the panels) and n^*, under three prior choices for α. The prior densities for α are denoted by the solid lines.

Survival analysis using Weibull DP mixtures

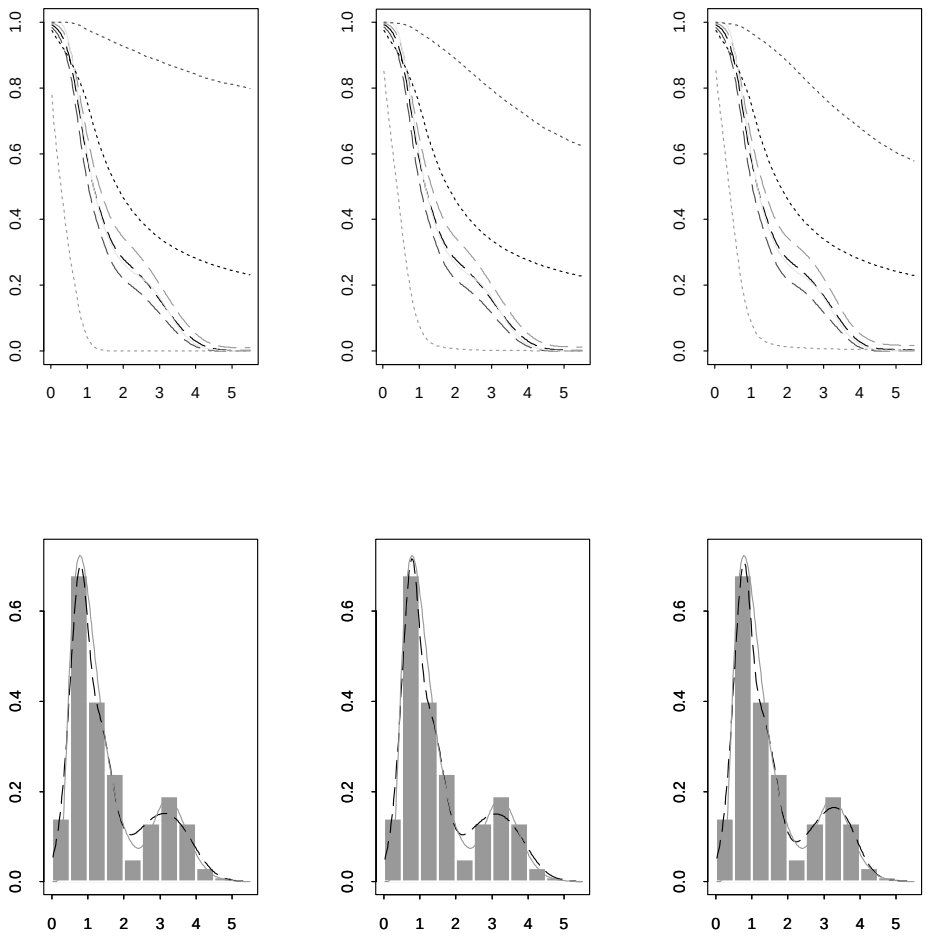


Figure 3.2: Simulated data. Inference, under three prior choices for α. The upper panels provide prior (dotted lines) and posterior (dashed lines) point and interval estimates for the survival function. The lower panels include the histogram of the data along with the posterior point estimate (dashed line) for the density function. In each graph, the solid line denotes the true curve.

Survival analysis using Weibull DP mixtures

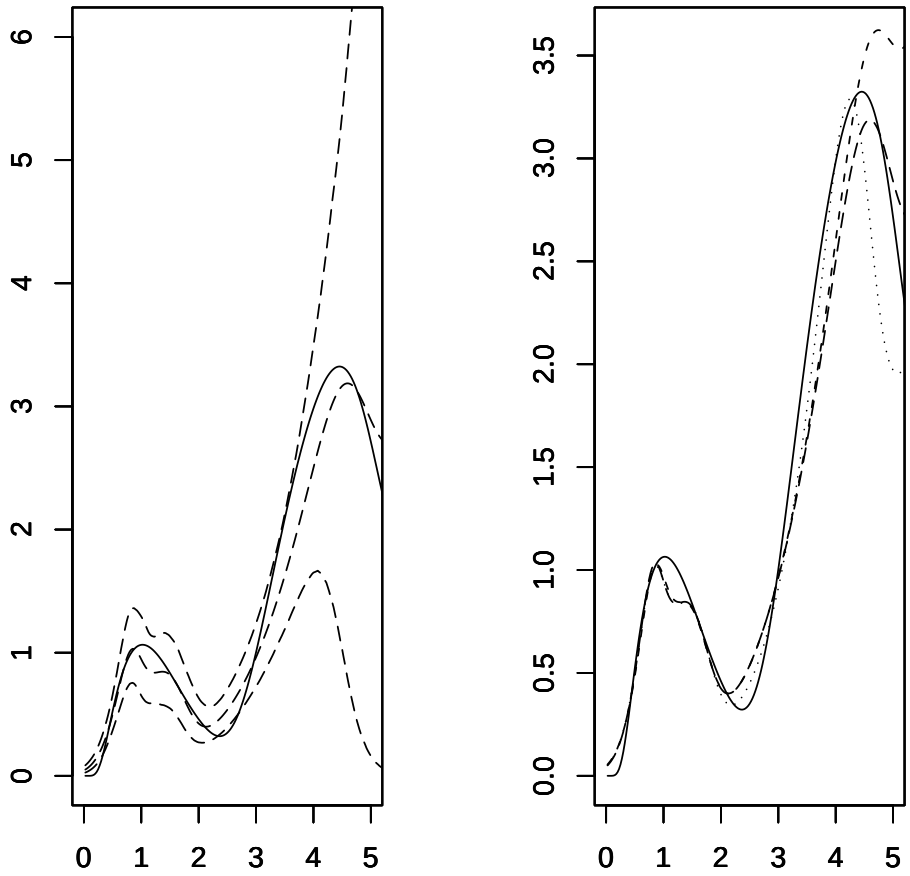


Figure 3.3: Simulated data. Posterior inference for the hazard function. Under a gamma(2,0.1) prior for α, the left panel provides point and interval estimates (dashed lines). The right panel compares point estimates under three priors for α, gamma(2,0.9) (smaller dashed line), gamma(2,0.1) (dashed line) and gamma(3,0.05) (dotted line). In each graph, the solid line denotes the true hazard function.

Survival analysis using Weibull DP mixtures

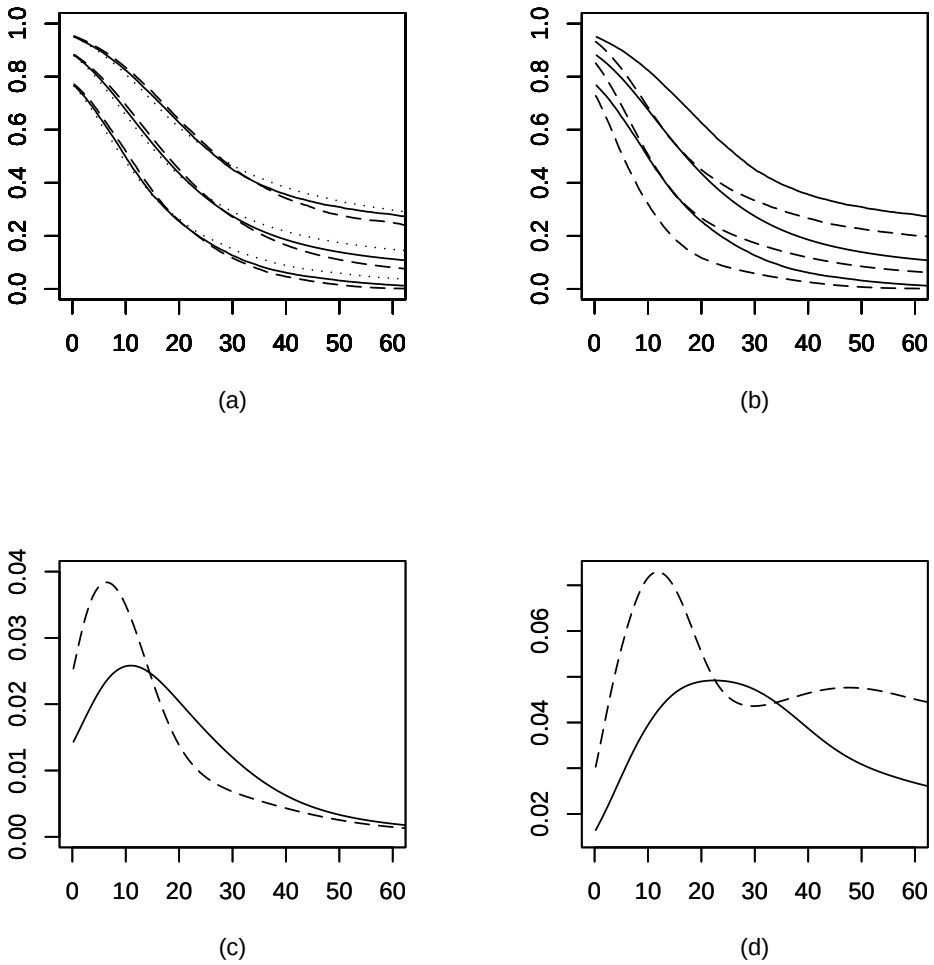


Figure 3.4: Data on remission times for leukemia patients. (a) Posterior point and interval estimates of the survival function for treatment A, under three different priors for α. Under the gamma(2,0.9) prior for α, Figures 4(b), 4(c) and 4(d) compare the survival functions (point and interval estimates), density functions and hazard functions (point estimates), respectively, for treatments A (solid lines) and B (dashed lines).

Survival analysis using Weibull DP mixtures

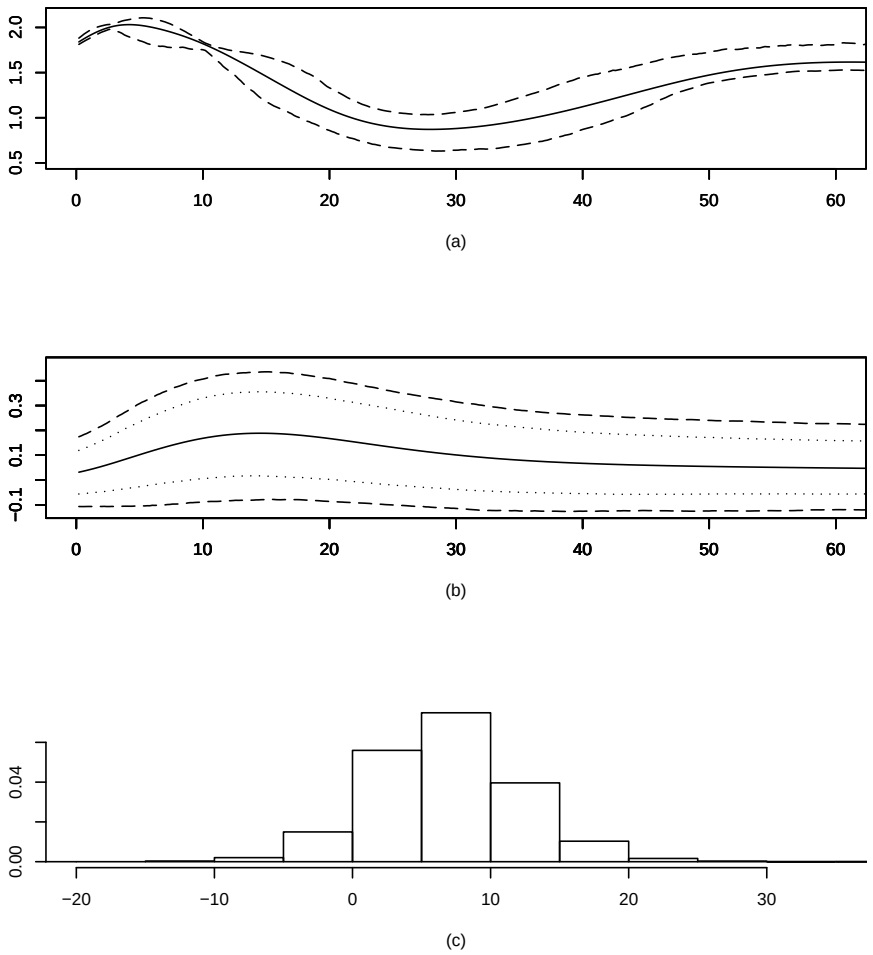


Figure 3.5: Data on remission times for leukemia patients. (a) Posterior point estimate (solid line) and 95% interval estimates (dashed lines) for $p(\lambda_B(t_0)/\lambda_A(t_0) \mid \text{data})$ (ratio of hazard functions). (b) Posterior point estimate (solid line), 80% interval estimates (dotted lines) and 95% interval estimates (dashed lines) for $p(F_B(t_0) - F_A(t_0) \mid \text{data})$ (difference of survival functions). (c) Histogram of draws from $p(\text{median}(A) - \text{median}(B) \mid \text{data})$ (difference of median survival times).

3.3 Curve fitting using Dirichlet process mixtures

- Two dominant trends in the Bayesian regression literature: seek increasingly flexible regression function models, and accompany these models with more comprehensive uncertainty quantification
- Typically, Bayesian nonparametric modelling focuses on either the regression function or the error distribution
- Bayesian nonparametric *implied conditional regression*: (West et al., 1994; Müller et al., 1996; Rodriguez et al., 2009; Taddy & Kottas, 2009a)
  → use flexible nonparametric mixture model for the joint distribution of response and covariates
  → inference for the conditional response distribution given covariates
- Both the response distribution and, implicitly, the regression relationship are modelled nonparametrically, thus providing a flexible framework for the general regression problem

Curve fitting using Dirichlet process mixtures

- Focus on univariate continuous response y (though extensions for categorical and/or multivariate responses also possible)

- DP mixture model for the joint density $f(y, \boldsymbol{x})$ of the response y and the vector of covariates $\boldsymbol{x}$:

$$f(y, \boldsymbol{x}) \equiv f(y, \boldsymbol{x}; G) = \int k(y, \boldsymbol{x}; \boldsymbol{\theta}) \mathrm{d}G(\boldsymbol{\theta}), \quad G \sim \mathrm{DP}(\alpha, G_0(\psi))$$

- For the mixture kernel $k(y, \boldsymbol{x}; \boldsymbol{\theta})$ use:
  $\rightarrow$ multivariate normal for (real-valued) continuous response and covariates
  $\rightarrow$ mixed continuous/discrete distribution to incorporate both categorical and continuous covariates
  $\rightarrow$ kernel component for y supported by $\mathbb{R}^+$ for problems in survival/reliability analysis

Curve fitting using Dirichlet process mixtures

- Again, introduce latent mixing parameters $\boldsymbol{\theta} = \{\boldsymbol{\theta}_i : i = 1, ..., n\}$ for each response/covariate observation $(y_i, \boldsymbol{x}_i)$, $i = 1, ..., n$ — **full posterior:**

$$p(G, \boldsymbol{\theta}, \alpha, \psi \mid \text{data}) = p(G \mid \boldsymbol{\theta}, \alpha, \psi) p(\boldsymbol{\theta}, \alpha, \psi \mid \text{data})$$

- $p(\boldsymbol{\theta}, \alpha, \psi \mid \text{data})$ is the posterior of the finite-dimensional parameter vector that results by marginalizing G over its DP prior
  $\rightarrow$ MCMC posterior simulation to sample from this marginal posterior

- $p(G \mid \boldsymbol{\theta}, \alpha, \psi)$ is a DP with precision parameter $\alpha + n$ and mean $(\alpha + n)^{-1} \left\{ \alpha G_0(\cdot; \psi) + \sum_{j=1}^{n^*} n_j \delta_{\boldsymbol{\theta}_j^*}(\cdot) \right\}$, where n^* is the number of distinct $\boldsymbol{\theta}_i$, and n_j is the size of the j-th distinct component
  $\rightarrow$ sample using the DP stick-breaking definition with a truncation approximation

- Alternatively, G can be truncated from the outset resulting in a finite mixture model that can be fitted with blocked Gibbs sampling

Curve fitting using Dirichlet process mixtures

- For any grid of values $(y_0, \boldsymbol{x}_0)$, obtain posterior samples for:
  $\rightarrow$ joint density $f(y_0, \boldsymbol{x}_0; G)$, marginal density $f(\boldsymbol{x}_0; G)$, and therefore, conditional density $f(y_0 \mid \boldsymbol{x}_0; G)$

  $\rightarrow$ conditional expectation $\mathrm{E}(y \mid \boldsymbol{x}_0; G)$, which, estimated over grid in $\boldsymbol{x}$, provides inference for the regression relationship
  $\rightarrow$ conditioning in $f(y_0 \mid \boldsymbol{x}_0; G)$ and/or $\mathrm{E}(y \mid \boldsymbol{x}_0; G)$ may involve only a portion of vector $\boldsymbol{x}$

- **Key features** of the modelling approach:
  $\rightarrow$ full and exact nonparametric inference (no need for asymptotics)
  $\rightarrow$ model for both non-linear regression curves **and** non-standard shapes for the conditional response density
  $\rightarrow$ model does not rely on additive regression formulations; it can uncover interactions between covariates that might influence the regression relationship

Data Example

- Simulated data set with a continuous response y, one continuous covariate x_c, and one binary categorical covariate x_d

  $\rightarrow$ x_{ci} ind. $\mathrm{N}(0,1)$

  $\rightarrow$ $x_{di} \mid x_{ci}$ ind. $\mathrm{Bernoulli}(\mathrm{probit}(x_{ci}))$

  $\rightarrow$ $y_i \mid x_{ci}, x_{di}$ ind. $\mathrm{N}(h(x_{ci}), \sigma_{x_{di}})$, with $\sigma_0 = 0.25$, $\sigma_1 = 0.5$, and

$$h(x_c) = 0.4x_c + 0.5\sin(2.7x_c) + 1.1(1 + x_c^2)^{-1}$$

  $\rightarrow$ two sample sizes: $n = 200$ and $n = 2000$

- DP mixture model with a mixed normal/Bernoulli kernel:

$$f(y, x_c, x_d; G) = \int \mathrm{N}_2(y, x_c; \boldsymbol{\mu}, \Sigma)\pi^{x_d}(1-\pi)^{1-x_d}\,\mathrm{d}G(\boldsymbol{\mu}, \Sigma, \pi),$$

  with $G \sim \mathrm{DP}(\alpha, G_0(\boldsymbol{\mu}, \Sigma, \pi) = \mathrm{N}_2(\boldsymbol{\mu}; \boldsymbol{m}, V) \times \mathrm{IWish}(\Sigma; \nu, S) \times \mathrm{Beta}(\pi; a, b))$

Curve fitting using Dirichlet process mixtures

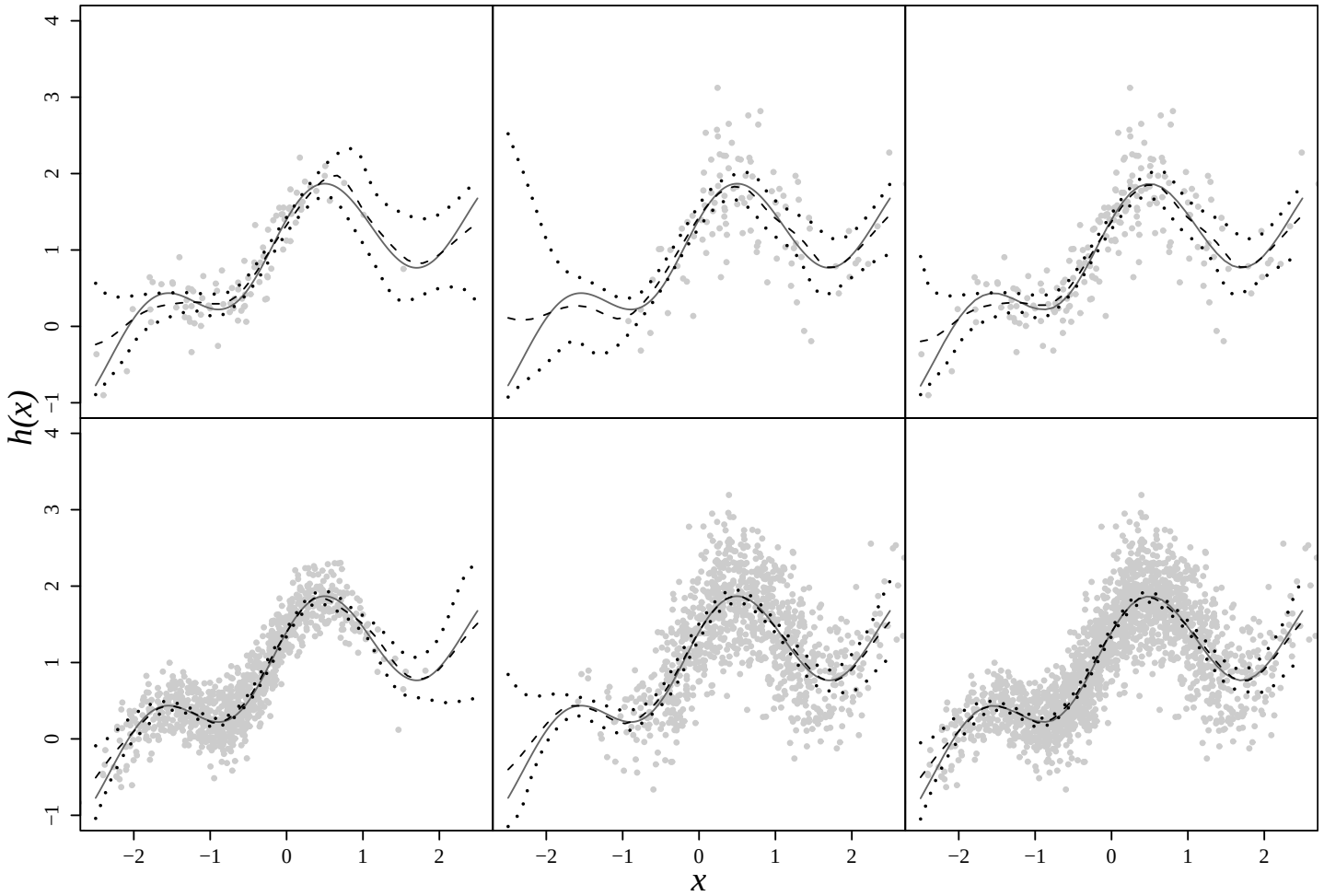


Figure 3.6: Posterior point and 90% interval estimates (dashed and dotted lines) for conditional response expectation $\mathrm{E}(y \mid x_c, x_d = 0; G)$ (left panels), $\mathrm{E}(y \mid x_c, x_d = 1; G)$ (middle panels), and $\mathrm{E}(y \mid x_c; G)$ (right panels). The corresponding data is plotted in grey for the sample of size $n = 200$ (top panels) and $n = 2000$ (bottom panels). The solid line denotes the true curve.

Curve fitting using Dirichlet process mixtures

- Model-based nonparametric approach to **quantile regression** (Taddy & Kottas, 2009a)
  $\rightarrow$ model joint density $f(y, \boldsymbol{x})$ of the response y and the M-variate vector of (continuous) covariates $\boldsymbol{x}$ with a DP mixture of normals:

$$f(y, \boldsymbol{x}; G) = \int \mathrm{N}_{M+1}(y, \boldsymbol{x}; \boldsymbol{\mu}, \Sigma) \mathrm{d}G(\boldsymbol{\mu}, \Sigma), \;\; G \sim \mathrm{DP}(\alpha, G_0)$$

  with $G_0(\boldsymbol{\mu}, \Sigma) = \mathrm{N}_{M+1}(\boldsymbol{\mu}; \boldsymbol{m}, V) \times \mathrm{IWish}(\Sigma; \nu, S)$

- For any grid of values $(y_0, \boldsymbol{x}_0)$, obtain posterior samples for:
  $\rightarrow$ conditional density $f(y_0 \mid \boldsymbol{x}_0; G)$ and conditional cdf $F(y_0 \mid \boldsymbol{x}_0; G)$
  $\rightarrow$ conditional quantile regression $q_p(\boldsymbol{x}_0; G)$, for any $0 < p < 1$

Curve fitting using Dirichlet process mixtures

- In regression settings, the covariates may have effect not only on the location of response distribution but also on its shape. Quantile regression quantifies relationship between a set of quantiles of response distribution and covariates, and thus, provides a more complete explanation of the response distribution in terms of available covariates

- Key features of the DP mixture modelling framework:
  $\rightarrow$ enables simultaneous inference for more than one quantile regression
  $\rightarrow$ allows flexible response distributions **and** non-linear quantile regression relationships
  $\rightarrow$ can be extended to handle partially observed responses (fully nonparametric Tobit quantile regression for econometrics data)

Data Example

- *Moral hazard* data on the relationship between shareholder concentration and several indices for managerial moral hazard in the form of expenditure with scope for private benefit (Yafeh & Yoshua, 2003)
  → data set includes a variety of variables describing 185 Japanese industrial chemical firms listed on the Tokyo stock exchange

  → response y: index *MH5*, consisting of general sales and administrative expenses deflated by sales
  → four-dimensional covariate vector $\boldsymbol{x}$: *Leverage* (ratio of debt to total assets); log(*Assets*); *Age* of the firm; and *TOPTEN* (the percent of ownership held by the ten largest shareholders)

Curve fitting using Dirichlet process mixtures

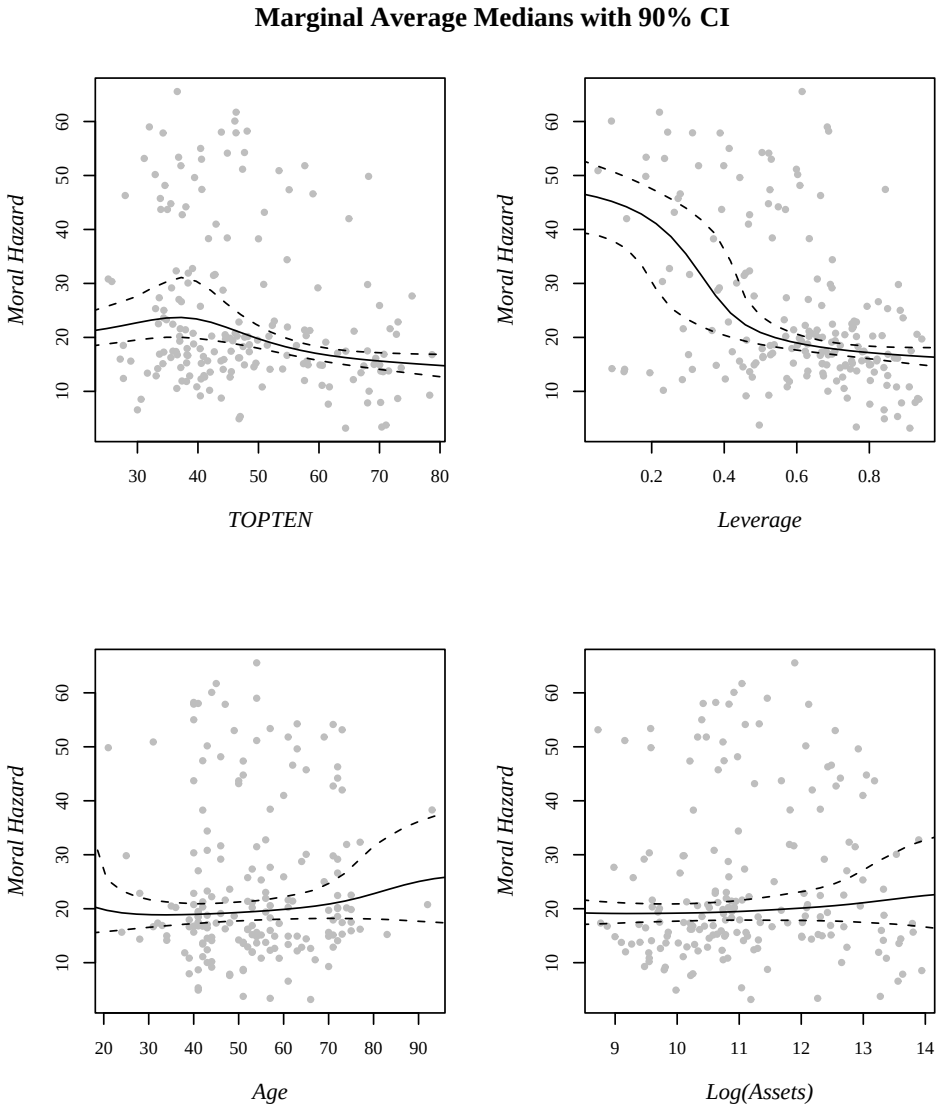


Figure 3.7: Posterior mean and 90% interval estimates for median regression for $MH5$ conditional on each individual covariate. Data scatterplots are shown in grey.

Curve fitting using Dirichlet process mixtures

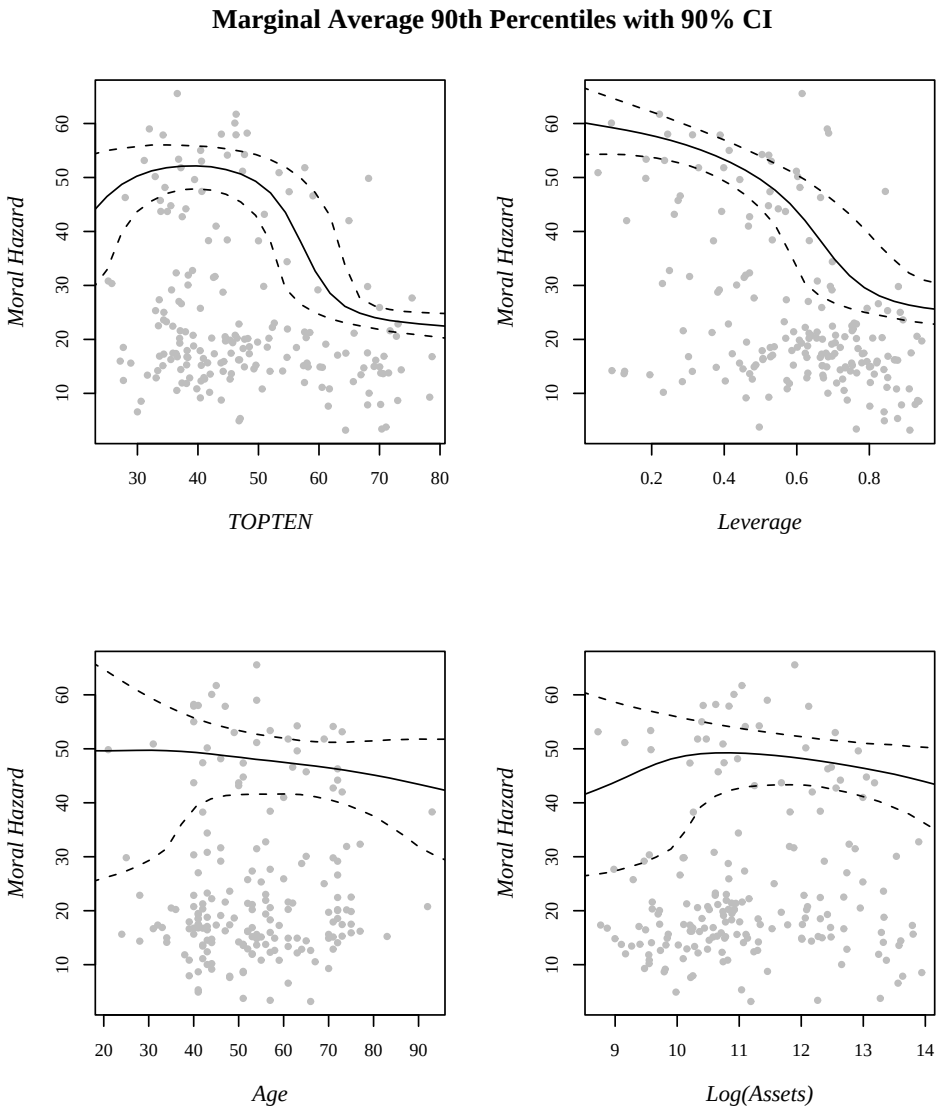


Figure 3.8: Posterior mean and 90% interval estimates for 90th percentile regression for $MH5$ conditional on each individual covariate. Data scatterplots are shown in grey.

Curve fitting using Dirichlet process mixtures

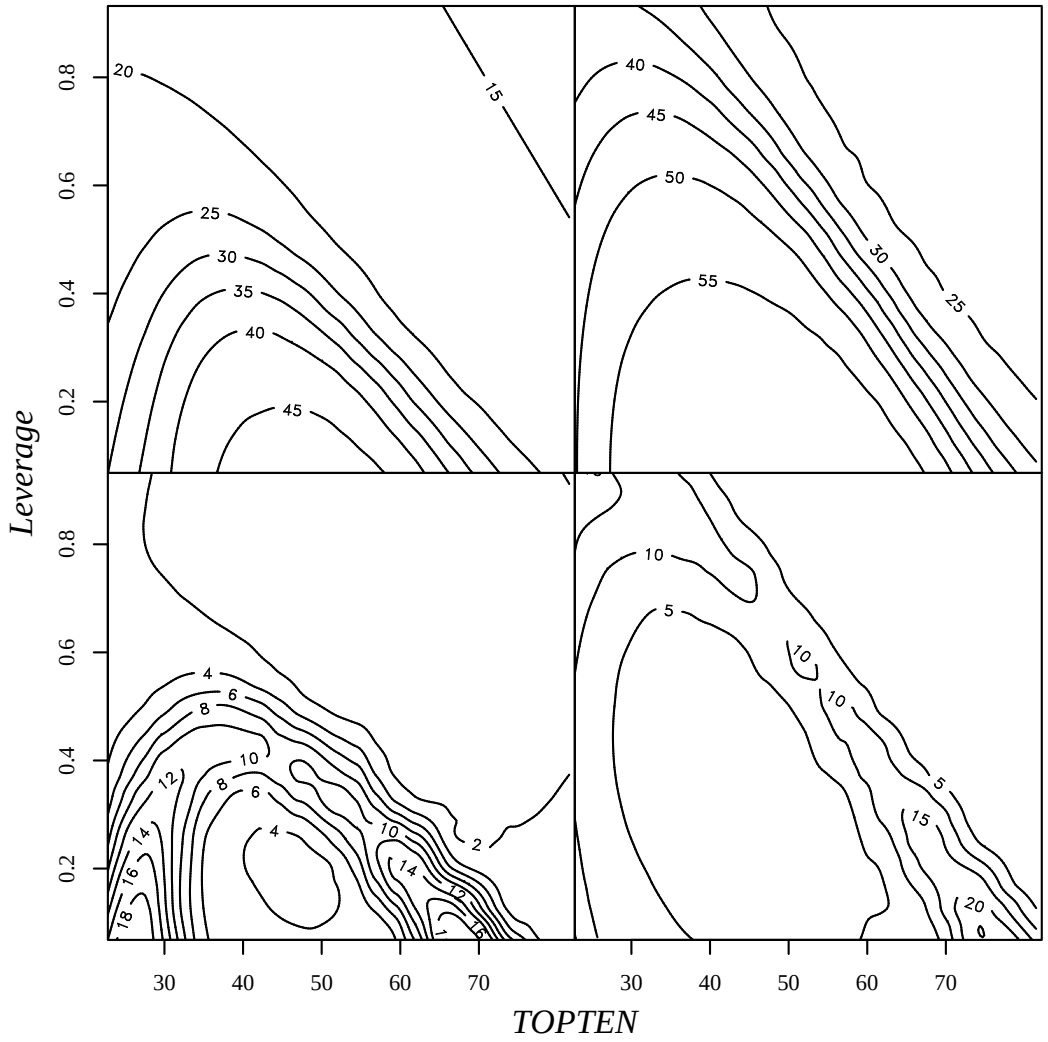


Figure 3.9: Posterior estimates of median surfaces (left column) and 90th percentile surfaces (right column) for $MH5$ conditional on *Leverage* and *TOPTEN*. The posterior mean is shown on the top row and the posterior interquartile range on the bottom.

Curve fitting using Dirichlet process mixtures

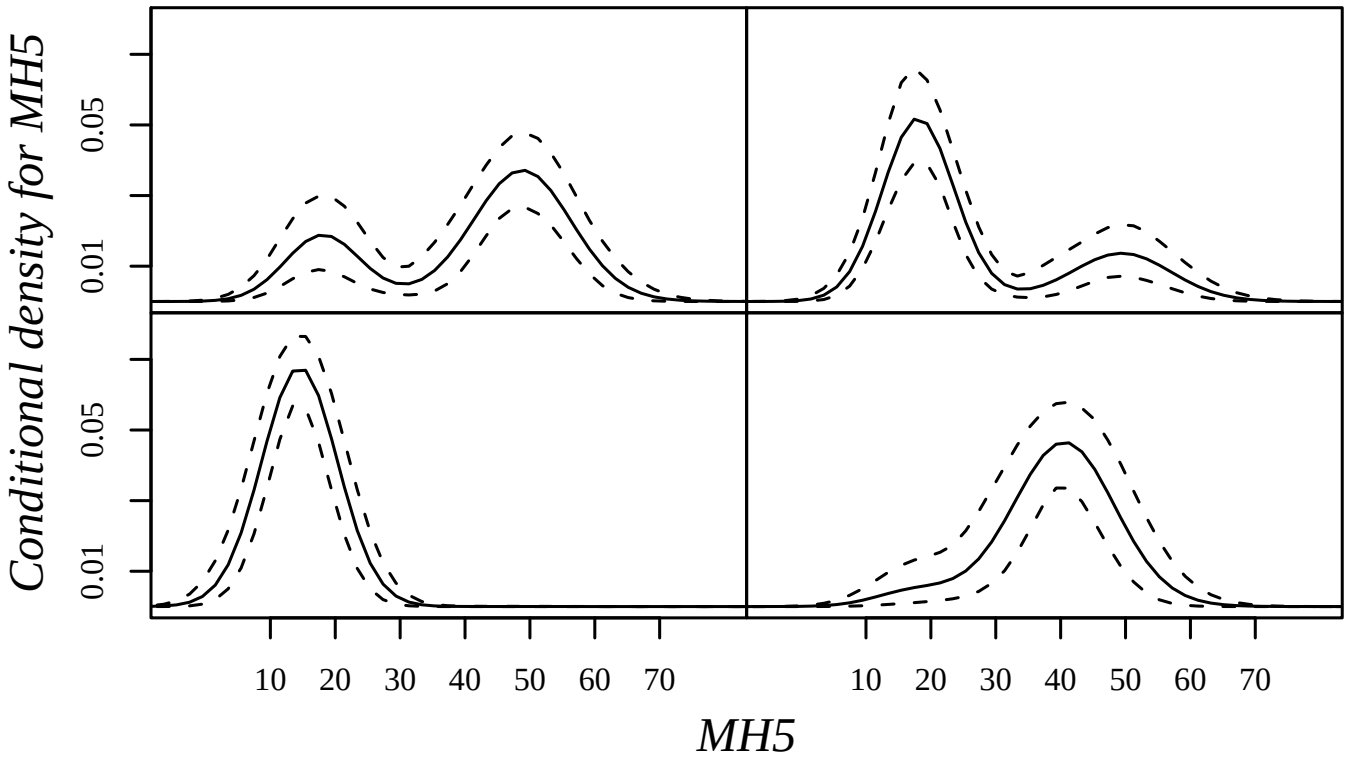


Figure 3.10: Posterior mean and 90% interval estimates for response densities $f(y \mid \boldsymbol{x}_0; G)$ conditional on four combinations of values $\boldsymbol{x}_0$ for the covariate vector ($TOPTEN$, $Leverage$, Age, $\log(Assets)$)

3.4 Modelling for multivariate ordinal data

- Values of k ordinal categorical variables $V_1, \ldots, V_k$ recorded for n subjects
  $\rightarrow C_j \geq 2$: number of categories for the j-th variable, $j = 1, \ldots, k$
  $\rightarrow n_{\ell_1 \cdots \ell_k}$: number of observations with $\boldsymbol{V} = (V_1, \ldots, V_k) = (\ell_1, \ldots, \ell_k)$
  $\rightarrow p_{\ell_1 \cdots \ell_k} = \Pr(V_1 = \ell_1, \ldots, V_k = \ell_k)$ is the classification probability for the $(\ell_1, \ldots, \ell_k)$ cell

- The data can be summarized in a k-dimensional contingency table with $C = \prod_{j=1}^{k} C_j$ cells, with frequencies $\{n_{\ell_1 \cdots \ell_k}\}$ constrained by $\sum_{\ell_1 \cdots \ell_k} n_{\ell_1 \cdots \ell_k} = n$

Modelling for multivariate ordinal data

- A possible modelling strategy (alternative to log-linear models) involves the introduction of k continuous latent variables $\boldsymbol{Z} = (Z_1, \ldots, Z_k)$ whose joint distribution yields the classification probabilities for the table cells, i.e.,

$$p_{\ell_1 \cdots \ell_k} = \Pr\left(\bigcap_{j=1}^{k} \left\{ \gamma_{j,\ell_j - 1} < Z_j \leq \gamma_{j,\ell_j} \right\} \right)$$

  for cutoff points $-\infty = \gamma_{j,0} < \gamma_{j,1} < \cdots < \gamma_{j,C_j - 1} < \gamma_{j,C_j} = \infty$, for each $j = 1, \ldots, k$ (e.g., Johnson & Albert, 1999)

- Common distributional assumption: $\boldsymbol{Z} \sim N_k(\boldsymbol{0}, \boldsymbol{S})$ (probit model)
  $\rightarrow \rho_{st} = \mathrm{Corr}(Z_s, Z_t) = 0$, $s \neq t$, implies independence of the corresponding categorical variables
  $\rightarrow$ coefficients ρ_{st}, $s \neq t$: *polychoric correlation coefficients* (traditionally used in the social sciences as a measure of association)

Modelling for multivariate ordinal data

- Richer modelling and inference based on normal DP mixtures for the latent variables $\boldsymbol{Z}_i$ associated with data vectors $\boldsymbol{V}_i$, $i = 1, ..., n$
- Model $\boldsymbol{Z}_i \mid G$ i.i.d. f, with $f(\cdot; G) = \int N_k(\cdot; \boldsymbol{m}, \boldsymbol{S}) \mathrm{d}G(\boldsymbol{m}, \boldsymbol{S})$, where

$$G \mid \alpha, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \boldsymbol{D} \sim \mathrm{DP}(\alpha, G_0(\boldsymbol{m}, \boldsymbol{S}) = N_k(\boldsymbol{m}|\boldsymbol{\lambda}, \boldsymbol{\Sigma}) \mathrm{IWish}_k(\boldsymbol{S}|\nu, \boldsymbol{D}))$$

- Advantages of the DP mixture modelling approach:
  $\rightarrow$ can accommodate essentially any pattern in k-dimensional contingency tables
  $\rightarrow$ allows local dependence structure to vary accross the contingency table
  $\rightarrow$ implementation does not require random cutoffs (so the complex updating mechanisms for cutoffs are not needed)

Data Example

- A Data Set of *Interrater Agreement*: data on the extent of scleral extension (extent to which a tumor has invaded the sclera or "white of the eye") as coded by two raters for each of $n = 885$ eyes
- The coding scheme uses five categories: 1 for "none or innermost layers"; 2 for "within sclera, but does not extend to scleral surface"; 3 for "extends to scleral surface"; 4 for "extrascleral extension without transection"; and 5 for "extrascleral extension with presumed residual tumor in the orbit"

- Results under the DP mixture model (and, for comparison, using also a probit model)
- The $(0.25, 0.5, 0.75)$ posterior percentiles for n^* are $(6, 7, 8)$ – in fact, $\Pr(n^* \geq 4 \mid \text{data}) = 1$

Table 3.1: For the interrater agreement data, observed cell relative frequencies (in bold) and posterior summaries for table cell probabilities (posterior mean and 95% central posterior intervals). Rows correspond to rater A and columns to rater B.

.3288 .3264 (.2940, .3586)	**.0836** .0872 (.0696, .1062)	**.0011** .0013 (.0002, .0041)	**.0011** .0020 (.0003, .0055)	**.0011** .0008 (.0, .0033)
.2102 .2136 (.1867, .2404)	**.2893** .2817 (.2524, .3112)	**.0079** 0.0080 (.0033, .0146)	**.0079** .0070 (.0022, .0143)	**.0034** .0030 (.0006, .0074)
.0023 .0021 (.0004, .0059)	**.0045** .0060 (.0021, .0118)	**.0** .0016 (.0004, .0037)	**.0023** .0023 (.0004, .0059)	**.0** .0009 (.0, .0030)
.0034 .0043 (.0012, .0094)	**.0113** .0101 (.0041, .0185)	**.0011** .0023 (.0004, .0058)	**.0158** .0142 (.0069, .0238)	**.0023** .0027 (.0006, .0066)
.0011 .0013 (.0001, .0044)	**.0079** .0071 (.0026, .0140)	**.0011** .0020 (.0003, .0054)	**.0090** .0084 (.0033, .0159)	**.0034** .0039 (.0011, .0090)

Modelling for multivariate ordinal data

- Posterior predictive distributions $p(\boldsymbol{Z}_0|\text{data})$ (Figure 3.11) – DP mixture version is based on the posterior predictive distribution for corresponding mixing parameter $(\boldsymbol{m}_0, \boldsymbol{S}_0)$
- Inference for the association between the ordinal variables:
  $\rightarrow$ e.g., Figure 3.11 shows posteriors for ρ_0, the correlation coefficient implied in $\boldsymbol{S}_0$
  $\rightarrow$ the probit model underestimates the association of the ordinal variables (as measured by ρ_0), since it fails to recognize clusters that are suggested by the data (as the DP mixture model reveals)
- Inference for log-odds ratios, $\psi_{ij} = \log p_{i,j} + \log p_{i+1,j+1} - \log p_{i,j+1} - \log p_{i+1,j}$ (Figure 3.12)
- Utility of mixture modelling for this data example – one of the clusters dominates the others, but identifying the other three is important; one of them corresponds to agreement for large values in the coding scheme; the other two indicate regions of the table where the two raters tend to agree less strongly

Modelling for multivariate ordinal data

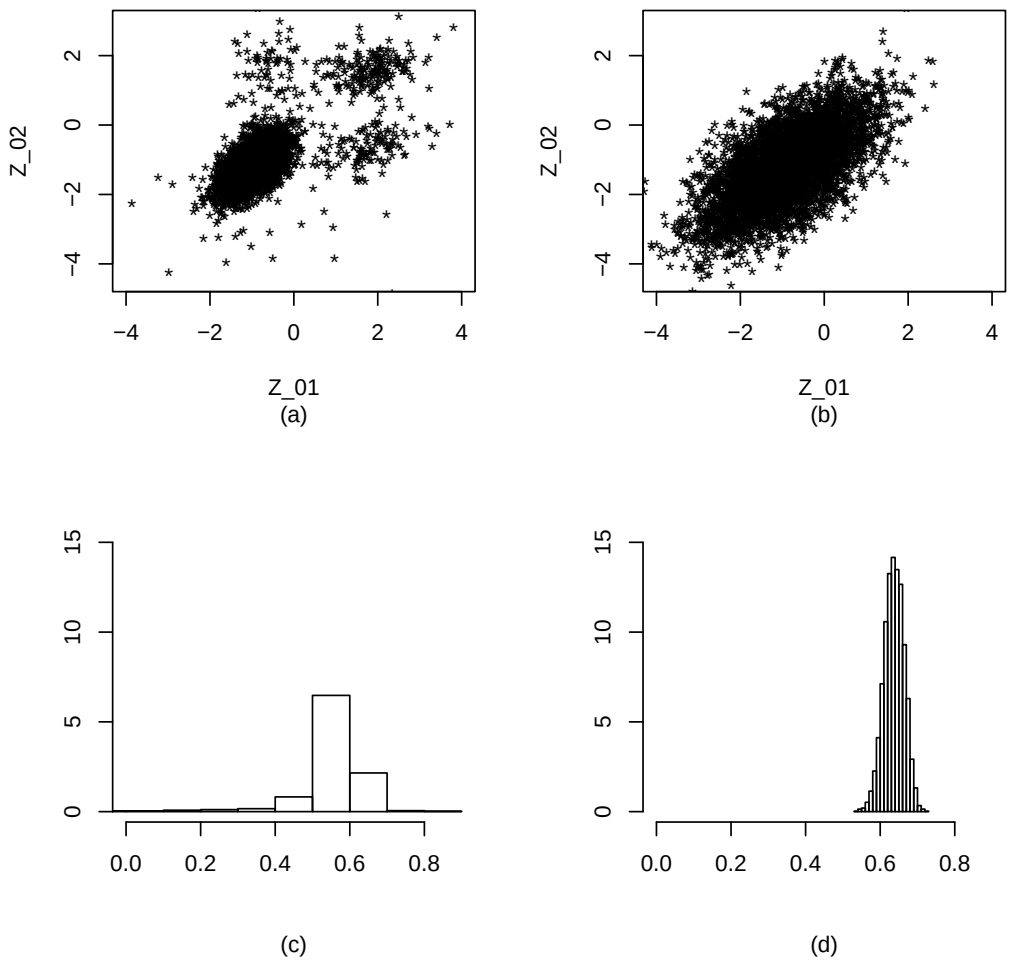


Figure 3.11: For the interrater agreement data, draws from $p(\boldsymbol{Z}_0|\text{data})$ and $p(\rho_0|\text{data})$ under the DP mixture model (panels (a) and (c), respectively) and the probit model (panels (b) and (d), respectively).

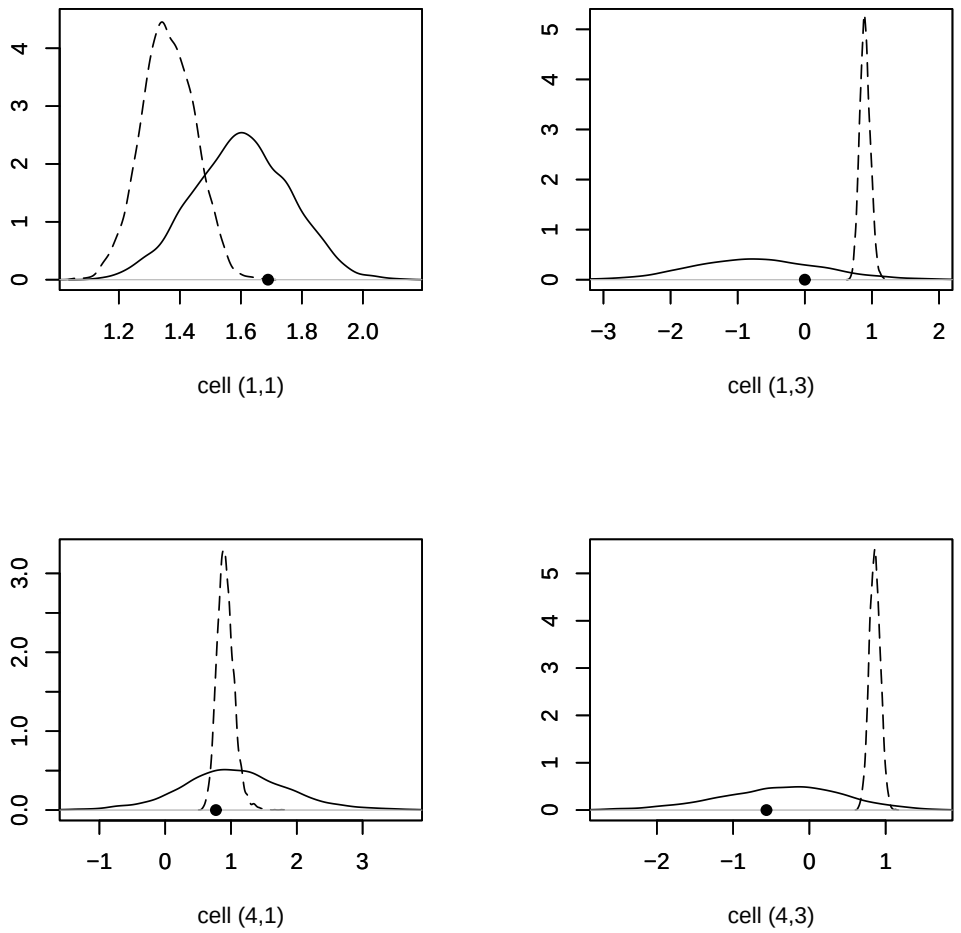


Figure 3.12: For the interrater agreement data, posteriors for four log-odds ratios under the DP mixture model (solid lines) and the probit model (dashed lines). The circles denote the corresponding empirical log-odds ratios.

3.5 Nonparametric inference for Poisson processes

- Point processes are stochastic process models for events that occur separated in time or space
- Applications of point process modeling in traffic engineering, software reliability, neurophysiology, weather modeling, forestry, ...
- Poisson processes, along with their extensions (Poisson cluster processes, marked Poisson processes, etc.), play an important role in the theory and applications of point processes (e.g., Kingman, 1993; Guttorp, 1995; Moller & Waagepetersen, 2004)

- Existing Bayesian nonparametric work based on gamma processes, weighted gamma processes, and Lévy processes (e.g., Lo & Weng, 1989; Kuo & Ghosh, 1997; Wolpert & Ickstadt, 1998; Gutiérrez-Peña & Nieto-Barajas, 2003; Ishwaran & James, 2004)

Nonparametric inference for Poisson processes

- For a point process over time, let $N(t)$ be the number of event occurrences in the time interval $(0, t]$ – the point process $\mathcal{N} = \{N(t) : t \geq 0\}$ is a non-homogeneous Poisson process (NHPP) if:
  $\rightarrow$ for any $t > s \geq 0$, $N(t) - N(s)$ follows a Poisson distribution with mean $\Lambda(t) - \Lambda(s)$, and
  $\rightarrow$ $\mathcal{N}$ has independent increments, i.e., for any $0 \leq t_1 < t_2 \leq t_3 < t_4$, $N(t_2) - N(t_1)$ and $N(t_4) - N(t_3)$ are independent random variables
- Λ is the mean measure (or cumulative intensity function) of the NHPP
- For any $t \in R^+$, $\Lambda(t) = \int_0^t \lambda(u)\mathrm{d}u$, where λ is the NHPP intensity function – λ is a non-negative and locally integrable function (i.e., $\int_B \lambda(u)\mathrm{d}u < \infty$, for all bounded $B \subset R^+$)

- So, from a modelling perspective, of interest for a NHPP is its intensity function

- Consider a NHPP observed over the time interval $(0, T]$ with events that occur at times $0 < t_1 < t_2 < ... < t_n \leq T$
- The likelihood for the NHPP intensity function λ is proportional to

$$\exp\{-\int_0^T \lambda(u)\mathrm{d}u\} \prod_{i=1}^{n} \lambda(t_i)$$

- **Key observation:** $f(t) = \lambda(t)/\gamma$, where $\gamma = \int_0^T \lambda(u)\mathrm{d}u$, is a density function on $(0, T)$
- Hence, (f, γ) provides an equivalent representation for λ, and so a nonparametric prior model for f, with a parametric prior for γ, will induce a semiparametric prior for λ — in fact, since γ only scales λ, it is f that determines the shape of the intensity function λ

- **Beta DP mixture model** for f (Kottas, 2006a)

$$f(t) \equiv f(t;G) = \int \text{be}(t;\mu,\tau)\text{d}G(\mu,\tau), \quad G \sim \text{DP}(\alpha, G_0)$$

  where $\text{be}(t;\mu,\tau)$ is the Beta density on $(0,T)$ with mean $\mu \in (0,T)$ and scale parameter $\tau > 0$, and $G_0(\mu,\tau) = \text{Unif}(\mu;0,T)$ inv-gamma $(\tau;c,\beta)$ with fixed shape parameter c and random scale parameter β

- Flexible density shapes through mixing of Betas (e.g., Diaconis & Ylvisaker, 1985) – Beta mixture model avoids edge effects (the main drawback of a normal DP mixture model in this setting)

- Full Bayesian model:

$$\exp(-\gamma)\gamma^n \left\{ \prod_{i=1}^{n} \int k(t_i;\mu_i,\tau_i)\text{d}G(\mu_i,\tau_i) \right\} p(\gamma)p(G \mid \alpha,\beta)p(\alpha)p(\beta)$$

  $\rightarrow$ DP prior structure $p(G \mid \alpha,\beta)p(\alpha)p(\beta)$ for G and its hyperparameters
  $\rightarrow$ reference prior for γ, $p(\gamma) \propto \gamma^{-1}$

Nonparametric inference for Poisson processes

- Letting $\boldsymbol{\theta} = \{(\mu_i, \tau_i) : i = 1, ..., n\}$, we have

$$p(\gamma, G, \boldsymbol{\theta}, \alpha, \beta|\text{data}) = p(\gamma|\text{data})p(G \mid \boldsymbol{\theta}, \alpha, \beta)p(\boldsymbol{\theta}, \alpha, \beta \mid \text{data})$$

  $\rightarrow$ $p(\gamma|\text{data})$ is a gamma$(n, 1)$ distribution

  $\rightarrow$ MCMC with Metropolis steps to sample from $p(\boldsymbol{\theta}, \alpha, \beta \mid \text{data})$

  $\rightarrow$ $p(G \mid \boldsymbol{\theta}, \alpha, \beta)$ is a DP with updated parameters – can be sampled using the DP constructive definition

- Full posterior inference for λ, Λ, and any other functional of the NHPP

- Application to neuronal data analysis (Kottas & Behseta, 2009) — extensions to inference for spatial NHPP intensities, using DP mixtures with bivariate Beta kernels (Kottas & Sansó, 2007)

- Extensions to semiparametric regression settings — models for **marked Poisson processes** (Taddy & Kottas, 2009b)

Data Illustrations

- Example for temporal NHPPs: data on the times of 191 explosions in mines, leading to coal-mining disasters with 10 or more men killed, over a time period of 40,550 days, from 15 March 1851 to 22 March 1962

- Prior specification for $\text{DP}(\alpha, G_0(\mu, \tau|\beta) = \text{Unif}(\mu|0, T)\text{IG}(\tau|2, \beta))$

  $\rightarrow$ gamma(a_α, b_α) prior for α – recall the role of α in controling the number n^* of distinct components in the DP mixture model

  $\rightarrow$ exponential prior for β – its mean can be specified using a prior guess at the range, R, of the event times t_i (e.g., $R = T$ is a natural default choice)

- Inference for the NHPP intensity under three prior choices: priors for β and α based on $R = T$, $\text{E}(n^*) \approx 7$; $R = T$, $\text{E}(n^*) \approx 15$; and $R = 1.5T$, $\text{E}(n^*) \approx 7$

- Examples for spatial NHPPs: two forestry data sets

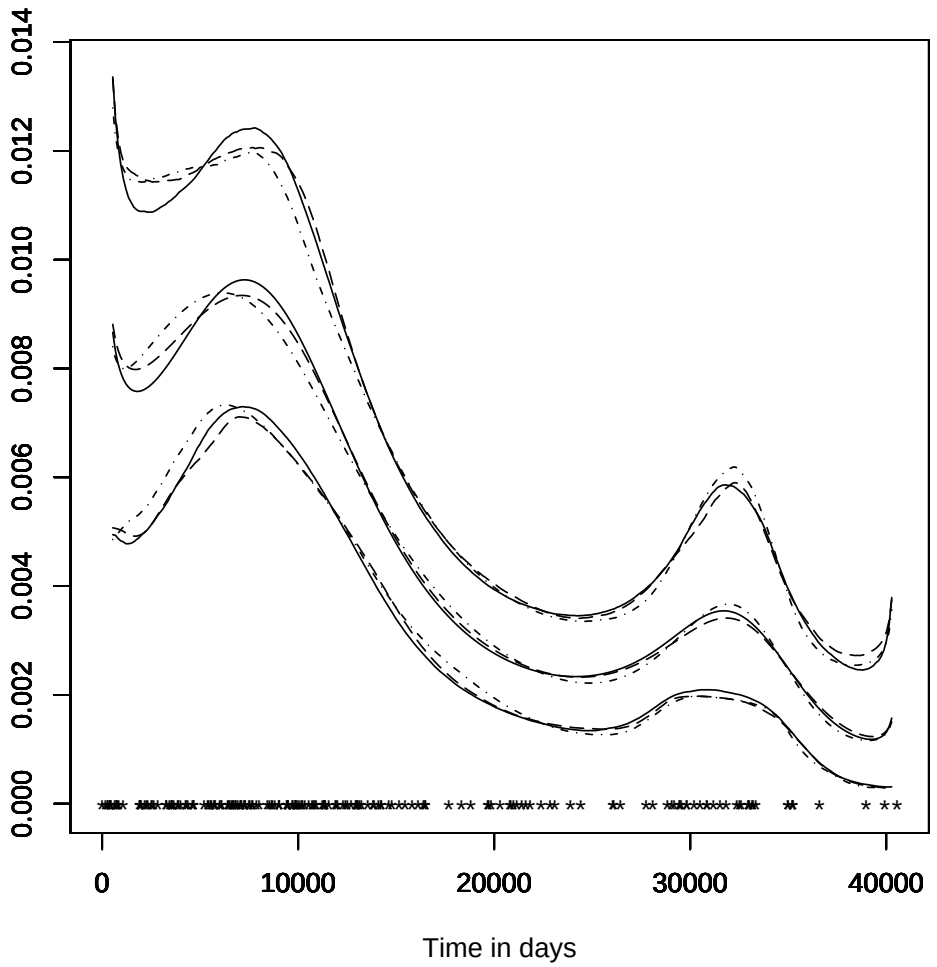


Figure 3.13: Coal-mining disasters data. Posterior point and 95% pointwise interval estimates for the intensity function under three prior settings. The observed times of the 191 explosions in mines are plotted on the horizontal axis.

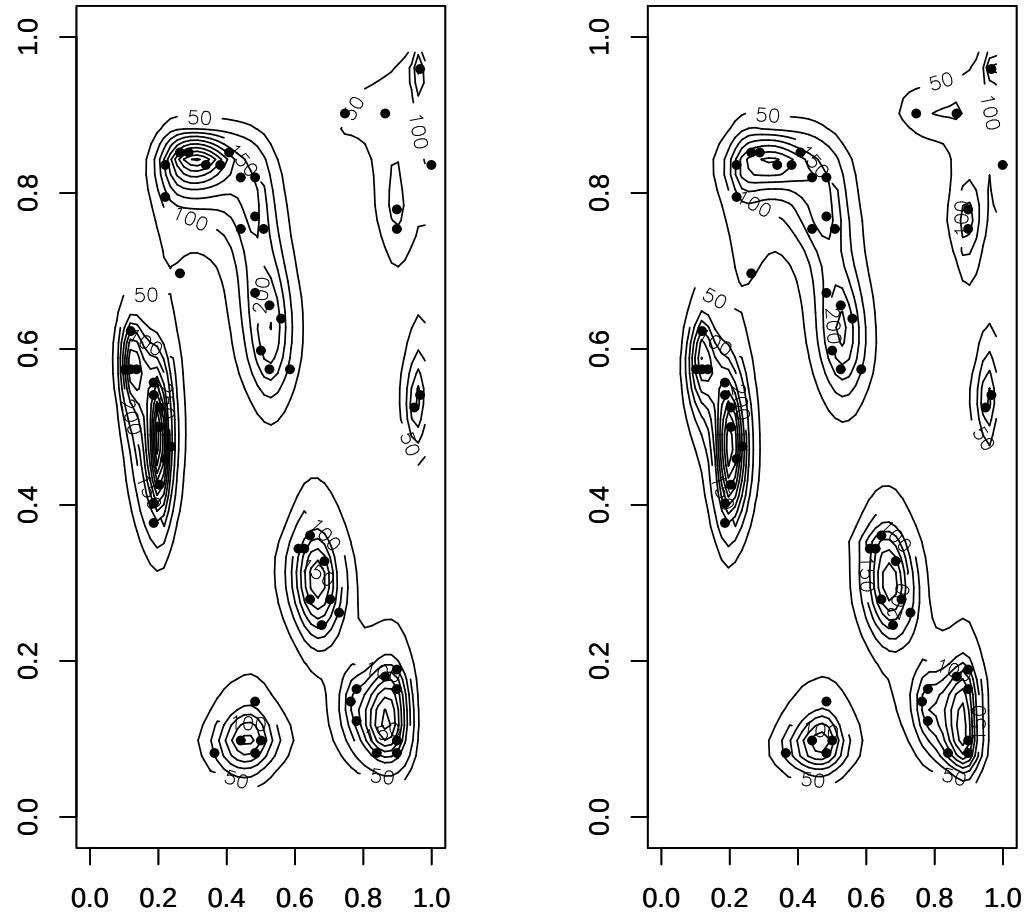


Figure 3.14: Redwood seedlings data. Contour plots of posterior mean intensity estimates under two different priors for α. The dots indicate the locations of the redwood seedlings.

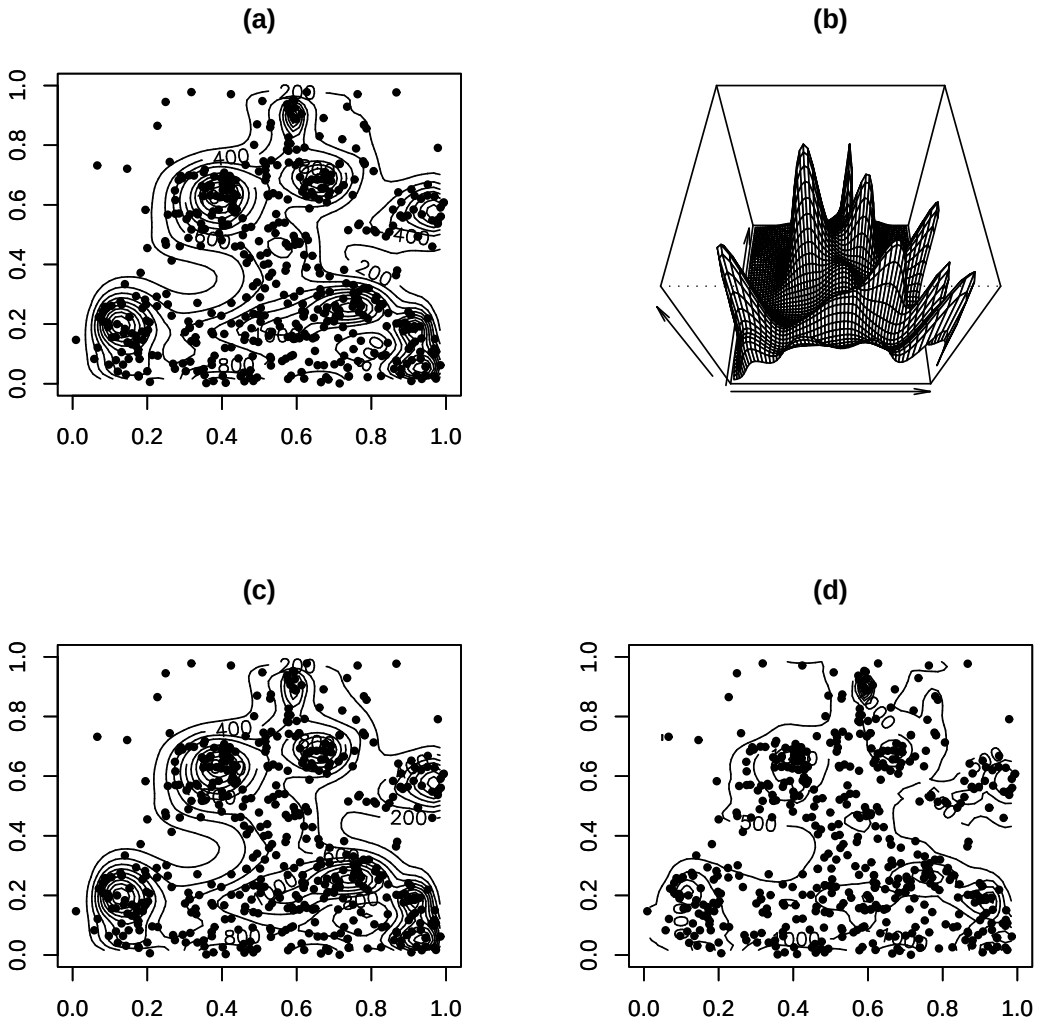


Figure 3.15: Maples data. Panels (a) and (b) include the posterior mean intensity estimate (contour and perspective plot, respectively). Panels (c) and (d) plot contour plots for the posterior median and interquartile range intensity estimates, respectively. The dots denote the locations of the maple trees.

3.6 Modelling for stochastically ordered distributions

- *Probability order* restrictions often appropriate/desirable when comparing two or more populations — different types of probability orders

- $\mathbb{R}$-valued random variables Y_1, Y_2 with distribution functions F_1, F_2
- **Stochastic order**: $Y_1 \leq_{st} Y_2$ (or $F_1 \leq_{st} F_2$) if, by definition,

$$F_1(u) \geq F_2(u),\ \forall u \in \mathbb{R} \ \Leftrightarrow\ \Pr(Y_1 > u) \leq \Pr(Y_2 > u),\ \forall u \in \mathbb{R}$$

$\rightarrow$ characterization: $Y_1 \leq_{st} Y_2$ if-f there exist r.v.s Y_1' and Y_2', defined on the same probability space, such that Y_1 and Y_2 have the same distribution with Y_1' and Y_2', and $\Pr(Y_1' \leq Y_2') = 1$

Modelling for stochastically ordered distributions

- Substantial literature on properties of distributions ordered according to one of these orders as well as on several other probability orders (Shaked & Shanthikumar, 1994) — also, extensive literature on classical estimation (typically, maximum likelihood estimation) and distribution-free testing for stochastic order, hazard rate order, and likelihood ratio order

- Arguments for forcing order restriction in the model:
  → order constraint of interest may not hold for the empirical distribution functions (especially for small or moderate sample sizes)
  → incorporating the order restriction can improve predictive accuracy
  → Bayesian framework attractive, since any order restriction in the prior model for the distributions is preserved to the posterior analysis

- **Bayesian nonparametric work:**
  stochastic and partial stochastic orders (Arjas & Gasbarra, 1996; Gelfand & Kottas, 2001; Hoff, 2003) – variability order (Kottas & Gelfand, 2001a) – stochastic precedence order (Chen & Dunson, 2004; Kottas, 2009)

A mixture modelling approach for stochastic order

- Focusing on two stochastically ordered distribution functions F_1 and F_2 (corresponding to distributions supported on $\mathbb{R}$), we seek nonparametric prior models over the space

$$\mathcal{P} = \{(F_1, F_2) : \ F_1 \leq_{st} F_2\}$$

- Constructive approach to building the restriction $F_1(u) \geq F_2(u)$, $u \in \mathbb{R}$, through latent distribution functions G_1 and G_2 (on $\mathbb{R}$) such that

$$F_1(u) = G_1(u), \quad F_2(u) = G_1(u)G_2(u)$$

(Note: with $\theta \sim G_1$ and independently $\delta \sim G_2$, F_1 and F_2 are the distributions of θ and $\max\{\theta, \delta\}$, respectively)

- Work with (independent) nonparametric priors for G_1 and G_2 to induce a prior over $\mathcal{P}' = \{(F_1, F_2) : F_1 = G_1, \ F_2 = G_1G_2\}$, and hence over $\mathcal{P}$

Modelling for stochastically ordered distributions

- How about using DP priors for G_1 and G_2
  $\rightarrow$ discreteness? simulation-based model fitting?

- Introduce DP mixing to overcome both difficulties

- **Key result:** for a parametric family of distributions $K(\cdot;\theta)$, $\theta \in (\underline{\theta}, \overline{\theta})$, strictly decreasing in θ, and H_1, H_2 two distribution functions on $(\underline{\theta}, \overline{\theta})$ with $H_1 \leq_{st} H_2$, defining

$$F(\cdot; H_i) = \int_{\underline{\theta}}^{\overline{\theta}} K(\cdot;\theta)\, \mathrm{d}H_i(\theta), \quad i = 1, 2$$

  we have $F(\cdot; H_1) \leq_{st} F(\cdot; H_2)$
  $\rightarrow$ result valid, e.g., for normal kernel mixing on the mean
  $\rightarrow$ add a dispersion parameter σ^2 to the model, to conclude that $F\left(\cdot; H_1, \sigma^2\right) \leq_{st} F\left(\cdot; H_2, \sigma^2\right)$ (semiparametric specification)

Modelling for stochastically ordered distributions

- Next, setting $H_1 = G_1$ and $H_2 = G_1 G_2$, we obtain the stochastically ordered DP mixture of normals model:

$$F_1(\cdot) \equiv F(\cdot; G_1, \sigma^2) = \int \mathrm{N}(\cdot; \theta, \sigma^2)\, \mathrm{d}G_1(\theta)$$

$$F_2(\cdot) \equiv F(\cdot; G_1, G_2, \sigma^2) = \iint \mathrm{N}(\cdot; \max\{\theta, \delta\}, \sigma^2)\, \mathrm{d}G_1(\theta)\mathrm{d}G_2(\delta)$$

  $\rightarrow$ with independent $\mathrm{DP}(\alpha_\ell, \mathrm{N}(\mu_\ell, \tau_\ell^2))$ priors for G_ℓ, $\ell = 1, 2$
  $\rightarrow$ an inverse gamma prior for σ^2, and priors for (a subset of) the hyperparameters $\boldsymbol{\psi} = \{\alpha_\ell, \mu_\ell, \tau_\ell^2 : \ell = 1, 2\}$

- Consider data $= \{y_{1i} : i = 1, ..., n_1;\ y_{2j} : j = 1, ..., n_2\}$ where the y_{1i} (given G_1, σ^2) are ind. from F_1 and the y_{2j} (given G_1, G_2, σ^2) are ind. from $F_2(\cdot)$

Modelling for stochastically ordered distributions

- Hierarchical formulation of the model:

$$
\begin{aligned}
y_{1i} \mid \theta_i, \sigma^2 &\overset{\text{ind.}}{\sim} \mathrm{N}(\theta_i, \sigma^2), \;\; i = 1, ..., n_1 \\
y_{2j} \mid \theta_{n_1+j}, \delta_j, \sigma^2 &\overset{\text{ind.}}{\sim} \mathrm{N}(\max\{\theta_{n_1+j}, \delta_j\}, \sigma^2), \;\; j = 1, ..., n_2 \\
\theta_i \mid G_1 &\overset{\text{i.i.d.}}{\sim} G_1, \;\; i = 1, ..., n_1 + n_2 \\
\delta_j \mid G_2 &\overset{\text{i.i.d.}}{\sim} G_2, \;\; j = 1, ..., n_2 \\
G_1, G_2 \mid \mu_1, \tau_1^2, \mu_2, \tau_2^2 &\sim \mathrm{DP}(\alpha_1, \mathrm{N}(\mu_1, \tau_1^2)) \times \mathrm{DP}(\alpha_2, \mathrm{N}(\mu_2, \tau_2^2))
\end{aligned}
$$

- Through the introduction of the additional mixing parameters θ_{n_1+j}, $j = 1, ..., n_2$, the first stage conditionally independent specification is retained after marginalizing G_1 and G_2 over their DP priors
- **Posterior inference:** simulation from the marginal posterior $p(\boldsymbol{\theta}, \boldsymbol{\delta}, \sigma^2, \boldsymbol{\psi} \mid \text{data})$, where $\boldsymbol{\theta} = \{\theta_i : i = 1, ..., n_1 + n_2\}$ and $\boldsymbol{\delta} = \{\delta_j : j = 1, ..., n_2\}$, enables estimation of posterior predictive densities

Modelling for stochastically ordered distributions

- More general inference requires the posteriors of G_1 and G_2:

  $$p(G_1, G_2, \boldsymbol{\theta}, \boldsymbol{\delta}, \sigma^2, \boldsymbol{\psi} \mid \text{data}) = p(G_1 \mid \boldsymbol{\theta}, \mu_1, \tau_1^2) p(G_2 \mid \boldsymbol{\delta}, \mu_2, \tau_2^2) p(\boldsymbol{\theta}, \boldsymbol{\delta}, \sigma^2, \boldsymbol{\psi} \mid \text{data})$$

  $\rightarrow$ where $p(G_1 \mid \boldsymbol{\theta}, \mu_1, \tau_1^2)$ denotes a DP distribution with precision parameter $\alpha_1 + n_1 + n_2$ and base distribution

  $$\frac{\alpha_1}{\alpha_1 + n_1 + n_2} \mathrm{N}(\cdot; \mu_1, \tau_1^2) + \frac{1}{\alpha_1 + n_1 + n_2} \sum_{i=1}^{n_1+n_2} \delta_{\theta_i}(\cdot)$$

  with analogous expressions for $p(G_2 \mid \boldsymbol{\delta}, \mu_2, \tau_2^2)$

- Sample from these two DPs using the DP stick-breaking representation with a truncation approximation

- Posterior samples for G_1, G_2 yield, for any set of grid points u, samples from the posterior of $F_1(u; G_1, \sigma^2)$ and $F_2(u; G_1, G_2, \sigma^2)$ (analogously for the mixture densities $f_1(u; G_1, \sigma^2)$ and $f_2(u; G_1, G_2, \sigma^2)$)

Applications to ROC analysis

- A commonly encountered task in epidemiologic research (both human and veterinary) involves the characterization of the discriminatory ability of a continuous diagnostic test
- In particular, *serologic scores* measure the concentration of antigen-specific antibodies in serum
- Commonly used continuous diagnostic measures result in an optical density value or a serum-to-positive ratio for an enzyme linked immunosorbent assay (ELISA) serological test — a relatively large serologic score is suggestive of disease or infection presence
- Data illustrations with commercially available ELISAs designed to detect antibodies to Johne's disease in dairy cows — Johne's disease is endemic throughout the US affecting multiple species of animals

Modelling for stochastically ordered distributions

- *Gold-standard* data setting: disease (infection) status is assumed known

- F_1 and F_2 are the distribution functions associated with serologic scores for the noninfected and infected populations, respectively — typically, F_1 and F_2 are modelled independently

- Incorporate stochastic order constraint $F_1 \leq_{st} F_2$ (Hanson et al., 2008)
- Biologically such a constraint is essentially always appropriate because serologic values for infected individuals tend to be larger than serologic values for noninfected individuals (provided the diagnostic test has reasonable discriminatory ability)
- Employ the stochastically ordered DP mixture model $F_1(\cdot) = F(\cdot; G_1, \sigma^2)$; $F_2(\cdot) = F(\cdot; G_1, G_2, \sigma^2)$

Modelling for stochastically ordered distributions

- Receiver operating characteristic (ROC) curve provides a commonly used graphical measure of the accuracy of the diagnostic test
- A cutoff value z can be used to dichotomize the serologic data into test positive (serologic score $> z$) or test negative (serologic score $< z$) categories
- ROC curve plots all possible pairs of true positive probability of infection $(1 - F_2(z))$ versus false positive probability $(1 - F_1(z))$ across all cutoff values z
- $\mathrm{ROC}(u) = 1 - F_2(F_1^{-1}(1-u))$, $u \in (0,1)$
- Area under the curve, $\mathrm{AUC} = \int_0^1 \mathrm{ROC}(u)\,\mathrm{d}u$ (probability that a randomly selected infected individual has a serologic score that is greater than that for a randomly selected noninfected individual)
- Posterior inference for $\mathrm{ROC}(\cdot)$ and AUC through the posteriors of $F(\cdot; G_1, \sigma^2)$ and $F(\cdot; G_1, G_2, \sigma^2)$

Modelling for stochastically ordered distributions

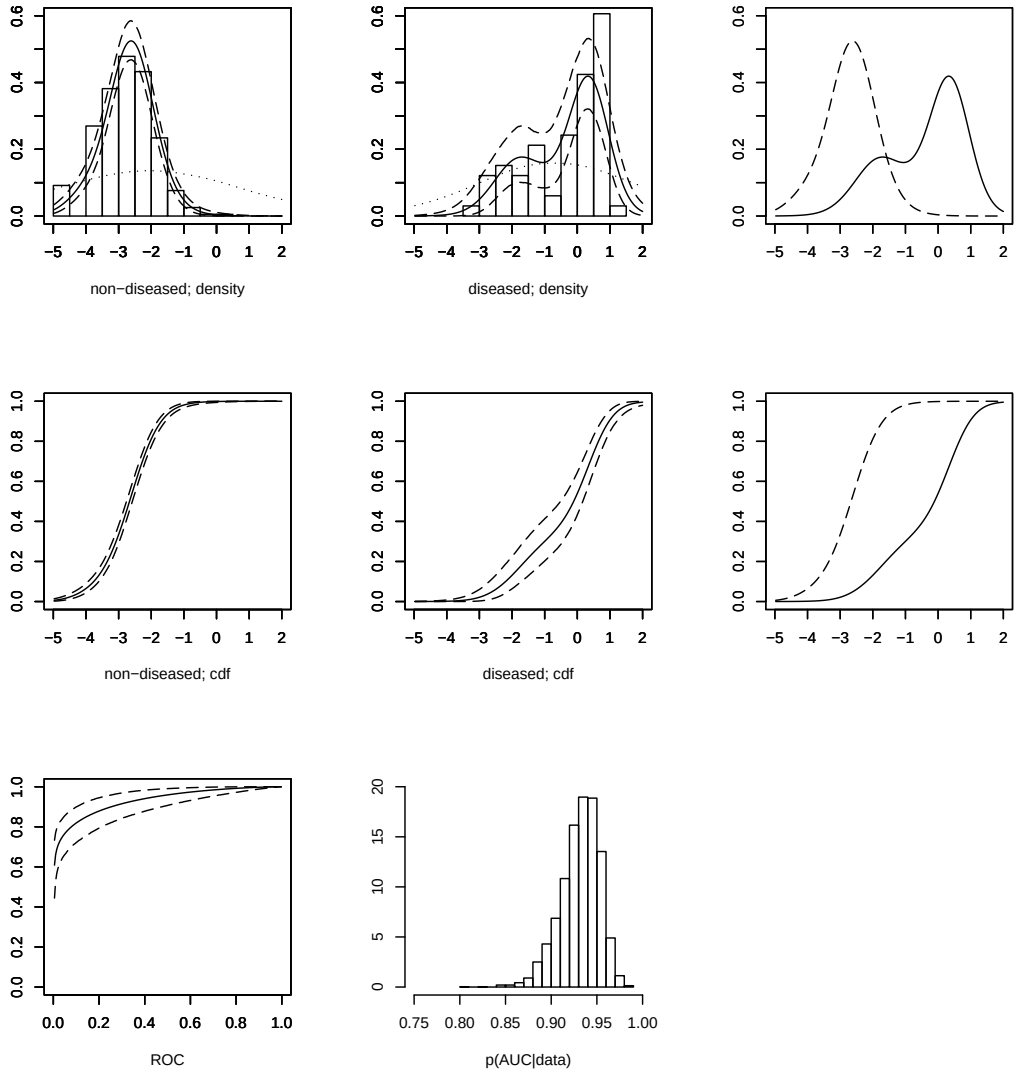


Figure 3.16: HerdChek ELISA test. Serologic scores for $n_1 = 393$ noninfected and $n_2 = 66$ infected cows.

Modelling for stochastically ordered distributions

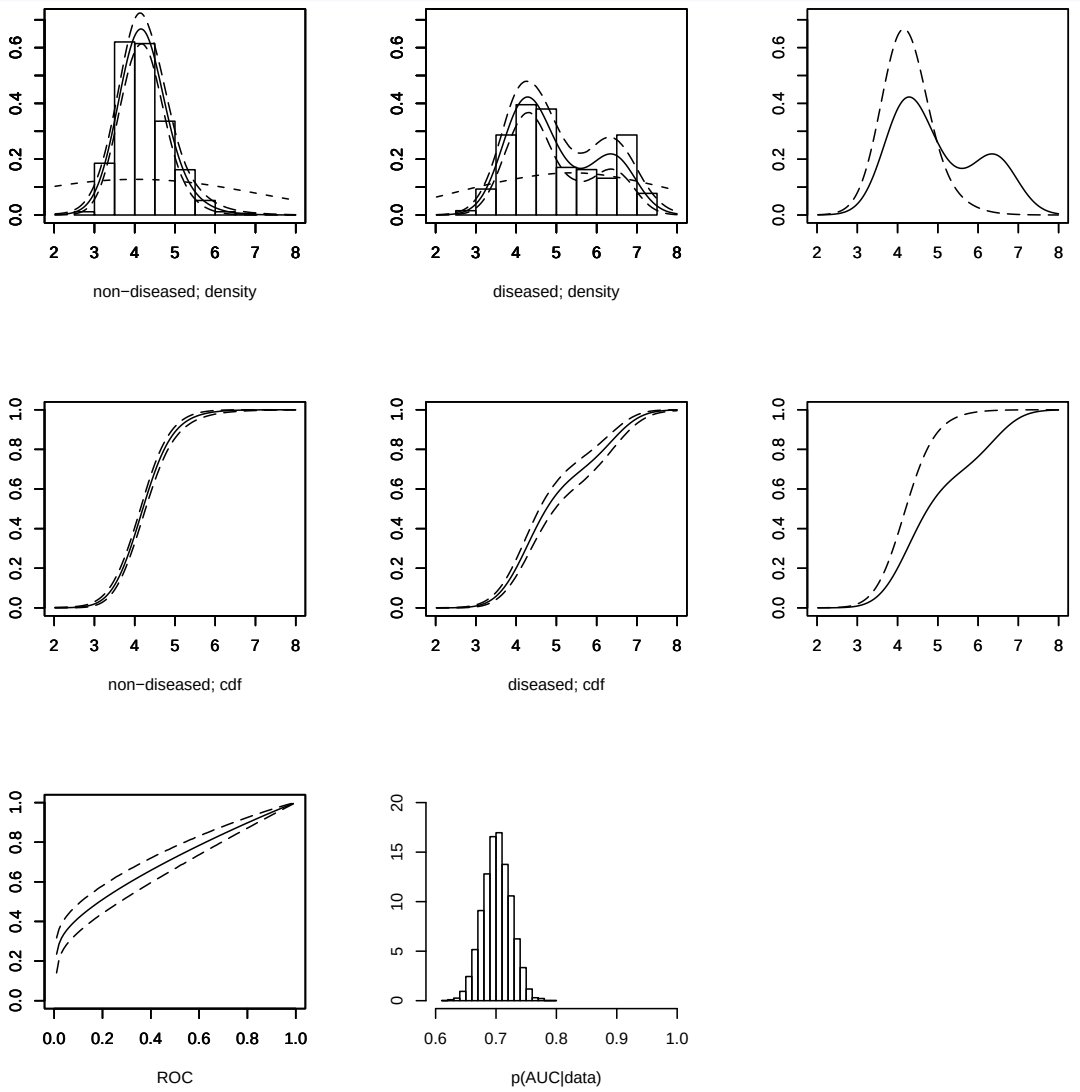


Figure 3.17: Institut Pourquier ELISA test. Scores for $n_1 = 345$ noninfected and $n_2 = 258$ infected cows.

Notes 4: Dependent Dirichlet process models

Outline

4.1 Dependent Dirichlet processes

- So far we have focused mostly on problems where a single distribution is assigned a nonparametric prior
- However, in many applications, the objective is modeling a collection of distributions $\mathcal{G} = \{G_{\mathbf{s}} : \mathbf{s} \in S\}$, where, for every $\mathbf{s} \in S$, $G_{\mathbf{s}}$ is a probability distribution — for example, S might be a time interval, a spatial region, or a covariate space
- Obvious options:
  $\rightarrow$ assume that the distribution is the same everywhere, e.g., $G_{\mathbf{s}} \equiv G \sim \mathrm{DP}(\alpha, G_0)$ for all $\mathbf{s}$. This is too restrictive
  $\rightarrow$ assume that the distributions are independent and identically distributed, e.g., $G_{\mathbf{s}} \sim \mathrm{DP}(\alpha, G_0)$ independently for each $\mathbf{s}$. This is wasteful
- We would like something in between

Dependent Dirichlet processes

- A similar dilemma arises in parametric models. Consider the hierarchical 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
- To obtain a setting that is *between* the two extremes above, the hierarchical model can be extended by assigning a prior to τ^2
- How can we generalize this idea to distributions?

Dependent Dirichlet processes

- A number of alternatives have been presented in the literature. We can classify them into three types of approaches:
  - Introducing dependence through the baseline distributions of conditionally independent nonparametric priors: for example, product of mixtures of DPs prior (refer to section 1.5). Simple but restrictive
  - Mixing of independent draws from a Dirichlet process:

$$G_{\mathbf{s}} = w_1(\mathbf{s})G_1^* + w_2(\mathbf{s})G_2^* + \ldots + w_p(\mathbf{s})G_p^*$$

    where $G_i^* \overset{ind.}{\sim} \mathrm{DP}(\alpha, G_0)$ and $\sum_{i=1}^{p} w_i(\mathbf{s}) = 1$ (e.g., Müller, Quintana & Rosner, 2004). Hard to extend to uncountable S
  - **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, all the models discussed here can be framed as DDPs

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

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

where $\delta_z(\cdot)$ denotes a point mass at z; the θ_ℓ are i.i.d. from G_0; and $\omega_1 = z_1$, $\omega_\ell = z_\ell \prod_{r=1}^{\ell-1}(1 - z_r)$, $\ell = 2, 3, ...$, with z_r i.i.d. $\mathrm{Beta}(1, \alpha)$

- 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

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

where $\theta_1(\mathbf{s}), \theta_2(\mathbf{s}), \ldots$ are sample paths from a (centering) stochastic process $G_{0,\mathbf{s}}$ defined on $\mathbf{s} \in S$, and $z_1(\mathbf{s}), z_2(\mathbf{s}), \ldots$ are sample paths from a stochastic process on S such that $z_\ell(\mathbf{s}) \sim \mathrm{Beta}(1, \alpha(\mathbf{s}))$

Dependent Dirichlet processes

- Key property: for any fixed $\mathbf{s}$, this construction yields a DP prior distribution for $G_{\mathbf{s}}$

- For any measurable set A, $G_{\mathbf{s}}(A)$ is a stochastic process with beta marginals. We can compute quantities like $\mathsf{Cov}(G_{\mathbf{s}}(A), G_{\mathbf{s}'}(A))$ (more on this later)

- As with Dirichlet processes, we usually employ the DDP prior to model the distribution of the parameters in a hierarchical model

- "Single-p" models $\Rightarrow$ The weights are assumed independent of $\mathbf{s}$, dependence over $\mathbf{s} \in S$ is due only to the dependence across atoms in the stick-breaking construction:

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

with $\omega_1 = z_1$, $\omega_\ell = z_\ell \prod_{r=1}^{\ell-1}(1 - z_r)$, $\ell = 2, 3, ...,$ with z_r i.i.d. Beta$(1, \alpha)$

  - Advantage $\Rightarrow$ Computation in DDP mixture models is simple, single-p DDP mixture models can be written as DP mixtures for an appropriate baseline process
  - Disadvantage $\Rightarrow$ It can be somewhat restrictive, for example, can never produce a collection of independent distributions, not even as a limiting case

Dependent Dirichlet processes

- In single-p models, for any measurable set A we have

$$\begin{aligned}
\mathsf{E}\{G_{\mathbf{s}}(A)\} &= \Pr\{\theta(\mathbf{s}) \in A\} \\
\mathsf{Var}\{G_{\mathbf{s}}(A)\} &= (1+\alpha)^{-1} \Pr\{\theta(\mathbf{s}) \in A\} [1 - \Pr\{\theta(\mathbf{s}) \in A\}] \\
\mathsf{Cov}\{G_{\mathbf{s}}(A), G_{\mathbf{s}'}(A)\} &= (1+\alpha)^{-1} (\Pr\{\theta(\mathbf{s}) \in A \cap \theta(\mathbf{s}') \in A\} \\
&\quad - \Pr\{\theta(\mathbf{s}) \in A\} \Pr\{\theta(\mathbf{s}') \in A\})
\end{aligned}$$

  where $\theta(\mathbf{s}) \sim G_{0,\mathbf{s}}$ (a sample path from the baseline process $G_{0,\mathbf{s}}$)

- Similarly, if $\eta(\mathbf{s}) \mid G_{\mathbf{s}} \sim G_{\mathbf{s}}$ and $G_{\mathbf{s}} \sim$ DDP, then a priori

$$\begin{aligned}
\mathsf{E}\{\eta(\mathbf{s})\} &= \mathsf{E}\{\theta(\mathbf{s})\} \\
\mathsf{Var}\{\eta(\mathbf{s})\} &= \frac{1}{1+\alpha} \mathsf{Var}\{\theta(\mathbf{s})\} \\
\mathsf{Cov}\{\eta(\mathbf{s}), \eta(\mathbf{s}')\} &= \frac{1}{1+\alpha} \mathsf{Cov}\{\theta(\mathbf{s}), \theta(\mathbf{s}')\}
\end{aligned}$$

where, again, all the expectations and covariances are computed under $G_{0,\mathbf{s}}$

Dependent Dirichlet processes

- We will discuss four classes of DDP models:
  - ANOVA DDP (De Iorio et al., 2004)
  - Hierarchical DPs (Teh. et al., 2006)
  - Nested DPs (Rodríguez, Dunson & Gelfand, 2008)
  - Spatial DPs (Gelfand, Kottas & MacEachern, 2005)
- However, this is by no means an exhaustive list: Order-depedent DDPs (Griffin & Steel, 2006), Generalized spatial DP (Duan, Guindani & Gelfand, 2007), Kernel stick-breaking process (Dunson & Park, 2007) ...

Example of DDP mixture modelling

- Application of a single-p DDP prior model to semiparametric quantile regression (Kottas & Krnjajić, 2009)

- More structured modelling formulation for quantile regression than the fully nonparametric approach discussed in Section 3.3

- Response observations y_i with covariate vectors x_i. Additive quantile regression formulation: $y_i = x_i'\beta + \varepsilon_i$, $i = 1, ..., n$
  $\rightarrow \varepsilon_i$ i.i.d. from an error distribution with p-th quantile equal to 0, i.e., $\int_{-\infty}^{0} f_p(\varepsilon)\mathrm{d}\varepsilon = p$

- **Objective**: develop flexible nonparametric prior models for the random error density $f_p(\cdot)$

- Key result: representation of non-increasing densities on $\mathbb{R}^+$ through scale uniform mixtures
- For any non-increasing density $f(\cdot)$ on $\mathbb{R}^+$ there exists a distribution function G, with support on $\mathbb{R}^+$, such that $f(t;G) = \int \theta^{-1} 1_{[0,\theta)}(t) \mathrm{d}G(\theta)$
- This result leads to a mixture representation for *any* unimodal density on the real line with p-th quantile (and mode) equal to zero, $\int\int k_p(\varepsilon;\sigma_1,\sigma_2)\mathrm{d}G_1(\sigma_1)\mathrm{d}G_2(\sigma_2)$, with G_1 and G_2 supported by $\mathbb{R}^+$, and

$$k_p(\varepsilon;\sigma_1,\sigma_2) = \frac{p}{\sigma_1} 1_{(-\sigma_1,0)}(\varepsilon) + \frac{(1-p)}{\sigma_2} 1_{[0,\sigma_2)}(\varepsilon),$$

with $0 < p < 1$, $\sigma_r > 0$, $r = 1, 2$

Dependent Dirichlet processes

- Assuming independent DP priors for G_1 and G_2, we obtain model:

$$f_p(\varepsilon; G_1, G_2) = \iint k_p(\varepsilon; \sigma_1, \sigma_2) \mathrm{d}G_1(\sigma_1) \mathrm{d}G_2(\sigma_2), \; G_r \sim \mathrm{DP}(\alpha_r, G_{r0}), r = 1, 2$$

  $\rightarrow$ the model can capture general forms of skewness and tail behavior

- The full hierarchical model:

$$\begin{aligned} y_i \mid \beta, \sigma_{1i}, \sigma_{2i} &\overset{ind}{\sim} k_p(y_i - x_i'\beta; \sigma_{1i}, \sigma_{2i}), \;\; i = 1, ..., n \\ \sigma_{ri} \mid G_r &\overset{iid}{\sim} G_r, \;\; r = 1, 2, \;\; i = 1, ..., n \\ G_r \mid \alpha_r, d_r &\sim \mathrm{DP}(\alpha_r, G_{r0} = \mathrm{IGamma}(c_r, d_r)), \;\; r = 1, 2 \end{aligned}$$

- **Quantile regression with dependent error densities:** relax the standard setting, under which the error density is the same for all x (so the response density changes with x only through the p-th quantile $x'\beta$)

Dependent Dirichlet processes

- To model nonparametrically quantile error distributions that change with covariates, we need a prior model for $f_{p,\mathcal{X}}(\cdot) = \{f_{p,x}(\cdot) : x \in \mathcal{X}\}$, where $\mathcal{X}$ is the covariate space and $\forall x, \int_{-\infty}^{0} f_{p,x}(\varepsilon)\mathrm{d}\varepsilon = p$
- To allow $f_p(\varepsilon; G_1, G_2)$ to change with x, the mixing distributions G_1, G_2 need to change with x — we need to replace G_r with a stochastic process $G_{r,\mathcal{X}}$ over $\mathcal{X}$
- DDP priors can be used for $G_{r,\mathcal{X}}$ — again, the idea is to use the constructive definition of the DP where now the point masses are i.i.d. realizations from a base stochastic process (say, a Gaussian process working with $\log(\sigma_{ri})$, $r = 1, 2$)
- A critical advantage of the DDP model is its flexibility in capturing different shapes for different covariate values (both observed and new covariate values)

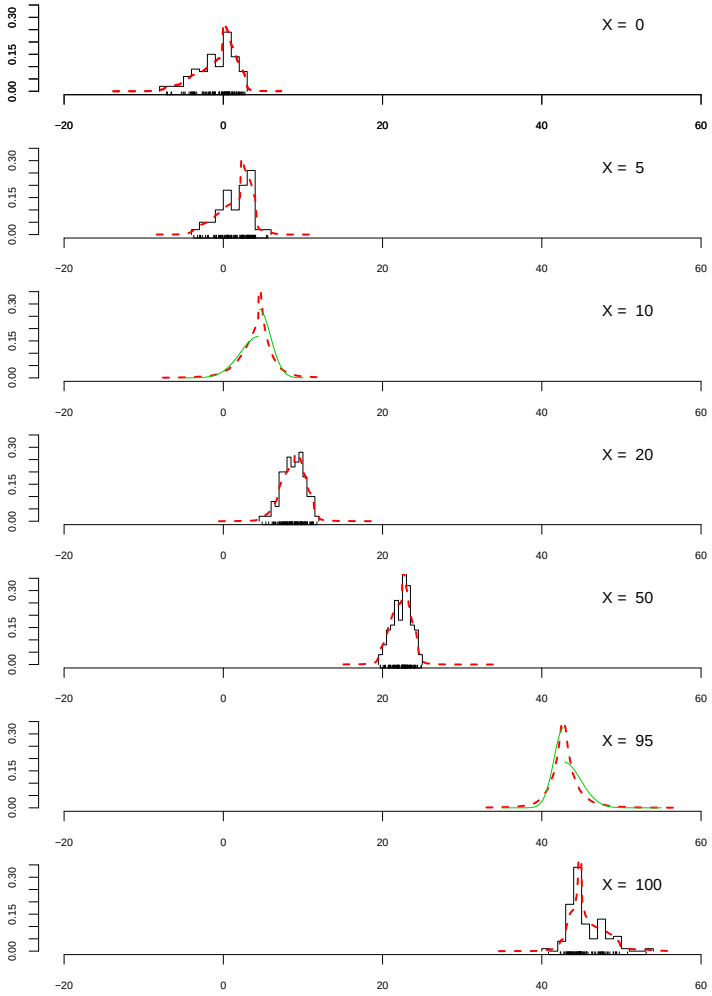


Figure 4.1: Simulation example. Posterior predictive densites under the DDP model (dashed lines) at the 5 observed covariate values (overlaid on corresponding data histograms), and at 2 new covariate values, $x = 10$ and $x = 95$ (overlaid on corresponding true densities denoted by solid lines).

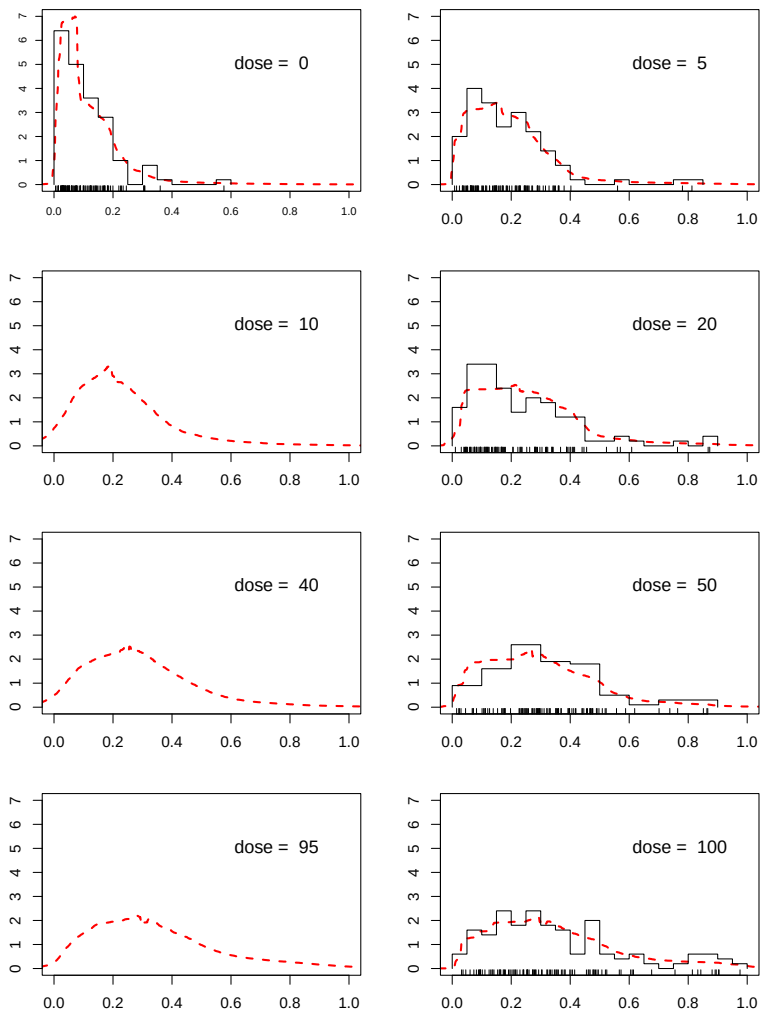


Figure 4.2: Comet assay data. Posterior predictive densites under the DDP model (dashed lines) at the 5 observed dose values, overlaid on corresponding data histograms, and at 3 new dose values (10, 40, and 95).

4.2 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. 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 $y_{s_1,s_2,k} \sim \int \mathrm{N}(y_{s_1,s_2,k}; \eta, \sigma^2) \mathrm{d}G_{s_1,s_2}(\eta)$ 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

$$m_h \sim G_0^m \qquad A_{h,s_1} \sim G_{0,s_1}^A \qquad B_{h,s_2} \sim G_{0,s_2}^B \qquad AB_{h,s_1,s_2} \sim G_{0,s_1,s_2}^{AB}$$

- Typically G_0^m, G_{0,s_1}^A, G_{0,s_2}^B and G_{0,s_1,s_2}^{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 mixture of ANOVA models where, at least in principle, there can be up to one different ANOVA for each observation:

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

  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: remember that the DP prior clusters observations. If a single component is used, we recover a parametric ANOVA model
- Rephrasing the model as a mixture simplifies computation: we can use a marginal sampler to fit the ANOVA-DDP model

Example with simulated data

- Data generated from two different mixture models, and we fit a one-way ANOVA-DDP with two levels
- Data in class one is generated from a single Gaussian distribution, while data in class 2 comes from a bimodal mixture of two Gaussian distributions
- 500 observations were generated in each group
- Function `LDDPdensity` in `DPpackage` can be used to fit the ANOVA-DDP model

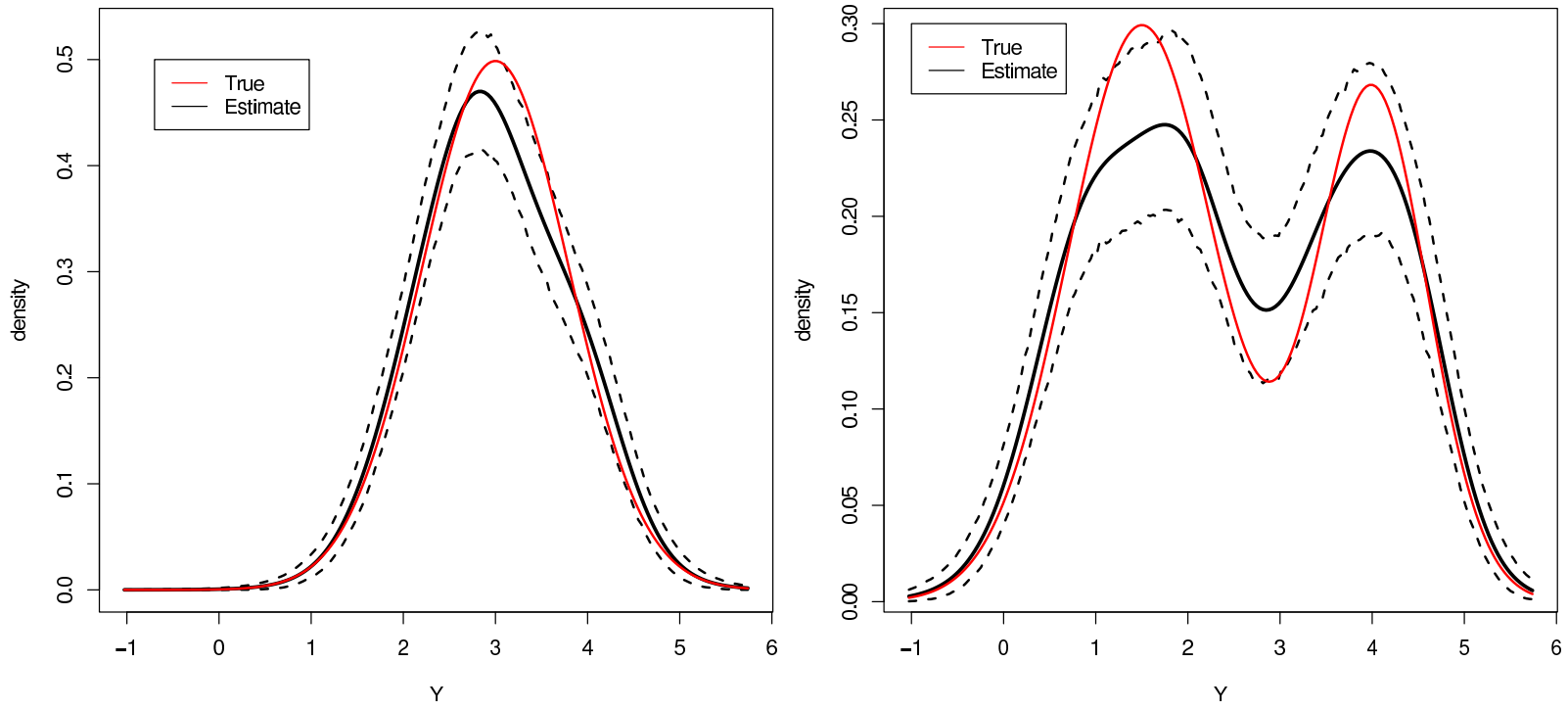


Figure 4.3: Density estimates for class one (left) and class two (right). The figure compares the true density (solid red line) with the ANOVA-DDP estimate (solid black line). Dashed lines correspond to pointwise credible bands.

4.3 Hierarchical Dirichlet processes

- We move now to a model for exchangeable collections of distributions
- Consider observations $y_{ij} \sim H_j$. For example, y_{ij} might correspond to the SAT score obtained by student $i = 1, \ldots, r_j$ in school $j = 1, \ldots, J$
- Hierarchical Dirichlet process (HDP) mixture models allow us to estimate H_j by identifying latent classes of students within each school. It also allows us to share classes across schools
- Let

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

- 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

- Since G_0 is drawn from a DP, it is almost surely discrete,

$$G_0 = \sum_{\ell=1}^{\infty} \omega_\ell \delta_{\phi_\ell}$$

- Therefore, when we draw the atoms for G_j we are forced to choose among $\phi_1, \phi_2, \ldots$, i.e., we can write G_j as:

$$G_j = \sum_{\ell=1}^{\infty} \pi_{\ell j} \delta_{\phi_\ell}$$

- Note that the weights assigned to the atoms *are not independent.* Intuitively, if ϕ_ℓ has a large associated weight ω_ℓ under G_0, then the weight $\pi_{\ell j}$ under G_j will likely be large for every j. Indeed,

$$\boldsymbol{\pi}_j = (\pi_{1j}, \pi_{2j}, \ldots) \sim \mathrm{DP}(\beta, \omega)$$

so that $\mathsf{E}(\pi_{\ell j}) = \omega_\ell$

- Note that, the HDP is a DDP, but not a single-p DDP (quite the contrary, you could call it a "single-atom" DDP).
- In spite of that, a simple MCMC sampler can be devised by composing two Pólya urns:

$$\theta_{ij}|\theta_{i-1,j},\ldots,\theta_{1,j},G_0 \sim \sum_{t=1}^{m_{j\cdot}} \frac{n_{jt\cdot}}{\alpha+i-1}\delta_{\psi_{jt}} + \frac{\alpha}{\alpha+i-1}G_0$$

$$\psi_{jt}|\psi_{11},\psi_{12},\ldots,\psi_{21},\ldots,\psi_{jt-1},H \sim \sum_{k=1}^{K} \frac{m_{\cdot k}}{m_{\cdot\cdot}+\beta}\delta_{\phi_k} + \frac{\beta}{m_{\cdot\cdot}+\beta}H$$

- The resulting algorithm is very similar to the marginal sampler for DP mixture models, but bookkeeping is a bit harder

4.4 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 together
- An example: assessment for quality of care in hospitals across the nation — we want to cluster states with similar distributions, and simultaneously cluster hospitals with similar outcomes
- Let y_{ij} be the percentage of patients in hospital $i = 1, \ldots, n_j$ within state $j = 1, \ldots, J$ who received the appropriate antibiotic on admission, then $y_{ij} \sim \int k(y_{ij}; \eta) \mathrm{d}G_j(\eta)$ where

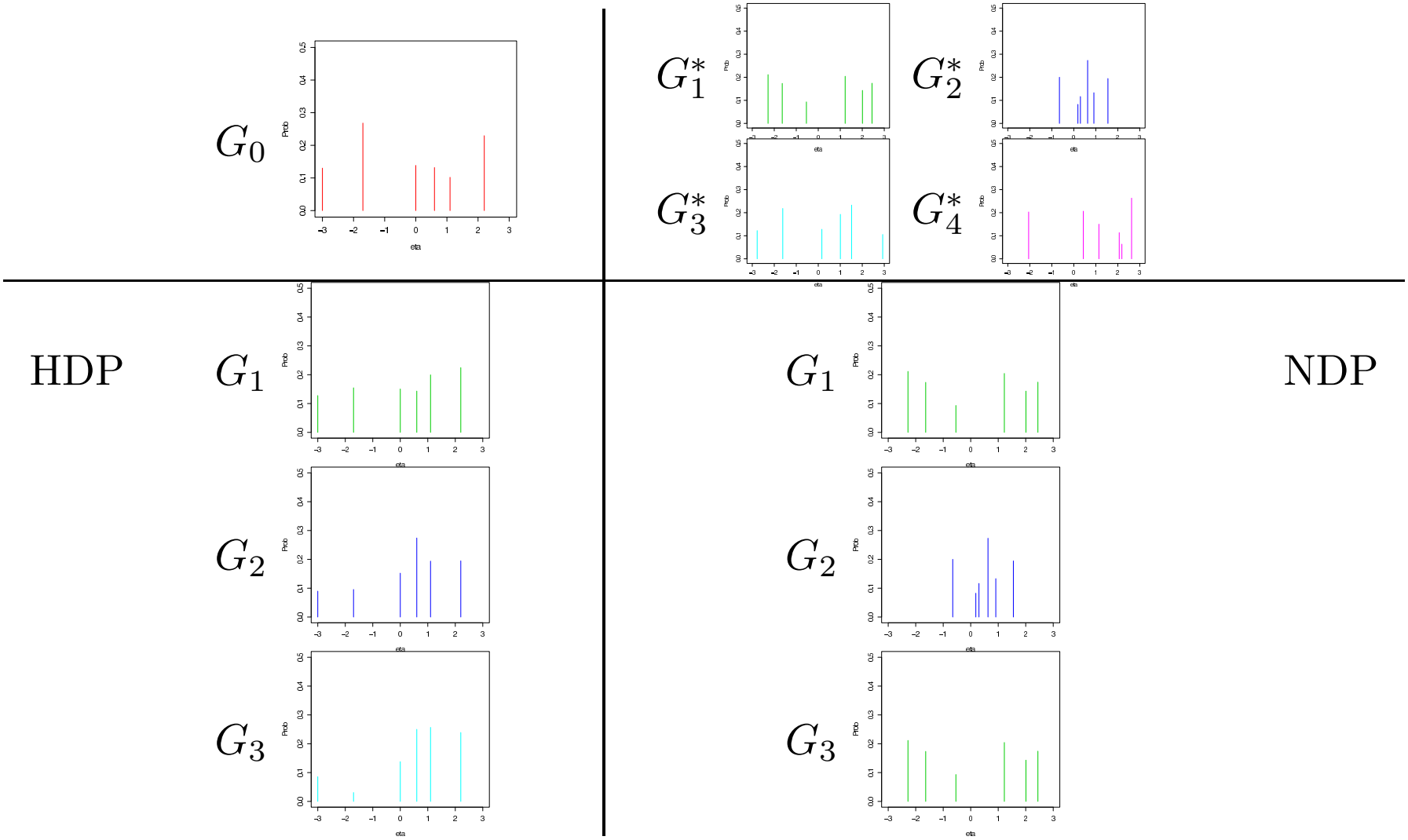
$$G_j \sim \sum_{k=1}^{K} \omega_k \delta_{G_k^*} \quad \text{and} \quad G_k^* = \sum_{\ell=1}^{\infty} \pi_{\ell k} \delta_{\theta_{\ell k}}$$

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 $\{G_1, \ldots, G_J\} \sim \text{DP}(\alpha, \text{DP}(\beta, H))$.
- Notationwise, the NDP resembles the HDP, but it is quite different
- The NDP is not a single-p DDP model
- 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
- A standard marginal sampler is not feasible in this problem — computation can be carried out using an extension of the blocked Gibbs sampler (see section 2.4.2)

Nested Dirichlet processes



4.5 Spatial Dirichlet process models

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

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

  - $\mu(s)$ mean process, e.g., $\mu(s) = x'(s)\boldsymbol{\beta}$
  - $\theta(s)$ 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. $\text{N}(0, \tau^2)$
- Induced model for observed sample (**point referenced spatial data**), $\mathbf{Y} = (Y(s_1), ..., Y(s_n))$, at sites $\boldsymbol{s}^{(n)} = (s_1, ..., s_n)$ in D

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

Spatial Dirichlet process models

- **Objective of Bayesian nonparametric modelling**: 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, $\mathbf{Y}_t = (Y_t(s_1), ..., Y_t(s_n))'$, $t = 1, ..., T$, 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
  $\rightarrow$ direct application of the methodology for spatial processes (when replications can be assumed approximately independent)
  $\rightarrow$ more generally, extension to **spatio-temporal modelling**, e.g., through dynamic spatial process modelling viewing $Y(s,t) \equiv Y_t(s)$ as a temporally evolving spatial process (Kottas, Duan & 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:

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

  random process over D centered at G_{0D} (notation: $G_D \sim \text{SDP}(\alpha, G_{0D})$)

- Key property: if

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

  then for any $\boldsymbol{s}^{(n)} = (s_1, ..., s_n)$, G_D induces $G^{(\boldsymbol{s}^{(n)})} \equiv G^{(n)}$, a random distribution for $(\theta(s_1), ..., \theta(s_n))$, and $G^{(n)} \sim \text{DP}(\alpha, G_0^{(n)})$, where $G_0^{(n)} \equiv G_0^{(\boldsymbol{s}^{(n)})}$ is n-variate normal (if G_{0D} is a Gaussian process)

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}
  $\rightarrow$ for instance, if G_{0D} produces a.s. continuous realizations, then $G^{(s)} - G^{(s')} \rightarrow 0$ a.s. as $||s - s'|| \rightarrow 0$
  $\rightarrow$ 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?
  $\rightarrow$ 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 & Gelfand, 2007)

Spatial Dirichlet process models

- Almost sure discreteness of realizations from G_D?
  $\rightarrow$ mix G_D against a pure error process $\mathcal{K}$ (i.i.d. variables $\epsilon(s)$ 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. $\text{N}(0, \tau^2)$
  (again, random process F is **non-Gaussian** and **nonstationary**)

- Adding covariates, the induced model at locations $\boldsymbol{s}^{(n)} = (s_1, ..., s_n)$,

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

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

Spatial Dirichlet process models

- Data: for $t = 1, ..., T$, response $\mathbf{Y}_t = (Y_t(s_1), ..., Y_t(s_n))'$ (with latent vector $\theta_t = (\theta_t(s_1), ..., \theta_t(s_n))'$), and matrix of covariate values 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 ($\rho_\phi(||\cdot||) = \exp(-\phi||\cdot||)$, $\phi > 0$, for the data examples)

- Bayesian model: (*conjugate* DP mixture model)

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

  with hyperpriors for $\boldsymbol{\beta}$, τ^2, α, σ^2, and ϕ

- Posterior inference using standard MCMC techniques for DP mixtures (refer to the second set of notes) — extensions to accommodate missing data — methods for prediction at new spatial locations

Data examples

- *Simulated data set*: generate data from a non-Gaussian process Z_D arising from a two-component mixture of independent Gaussian processes with constant means μ_k and covariance functions $\sigma^2 \exp(-\phi_k ||s - s'||)$, $k = 1, 2$ (processes 1 and 2 are sampled with probabilities q and $1 - q$)
- Generate $T = 75$ replications (at the $n = 39$ sites given in Figure 4.4) from the process $Z(s) + e(s)$, where $\sigma = 0.5$, $\phi_1 = \phi_2 = 0.0025$, $\mu_1 = -2$, $\mu_2 = 2$, $q = 0.75$, and $e(s)$ is a pure error process with variance $\tau^2 = 0.5$
- New sites for spatial prediction (denoted by "$*$" in Figure 4.4)

- Comparison with a Gaussian process mixture model,
  $\theta_t \mid \sigma^2, \phi \overset{i.i.d.}{\sim} G_0^{(n)} = \mathrm{N}_n(\cdot \mid 0_n, \sigma^2 H_n(\phi))$, $t = 1, ..., T$
  (limiting case, as $\alpha \to \infty$, of the spatial DP mixture model)

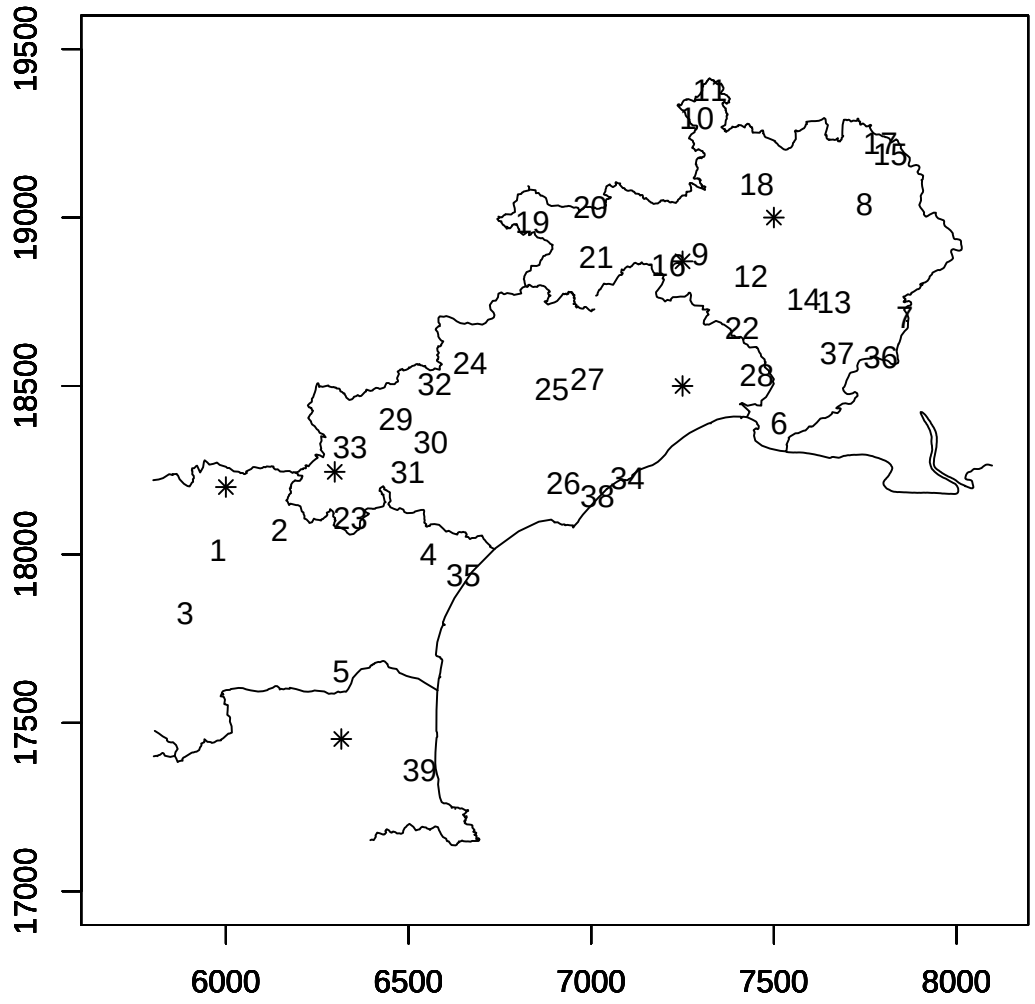


Figure 4.4: Geographic map of the Languedoc-Roussillon region in southern France

Spatial Dirichlet process models

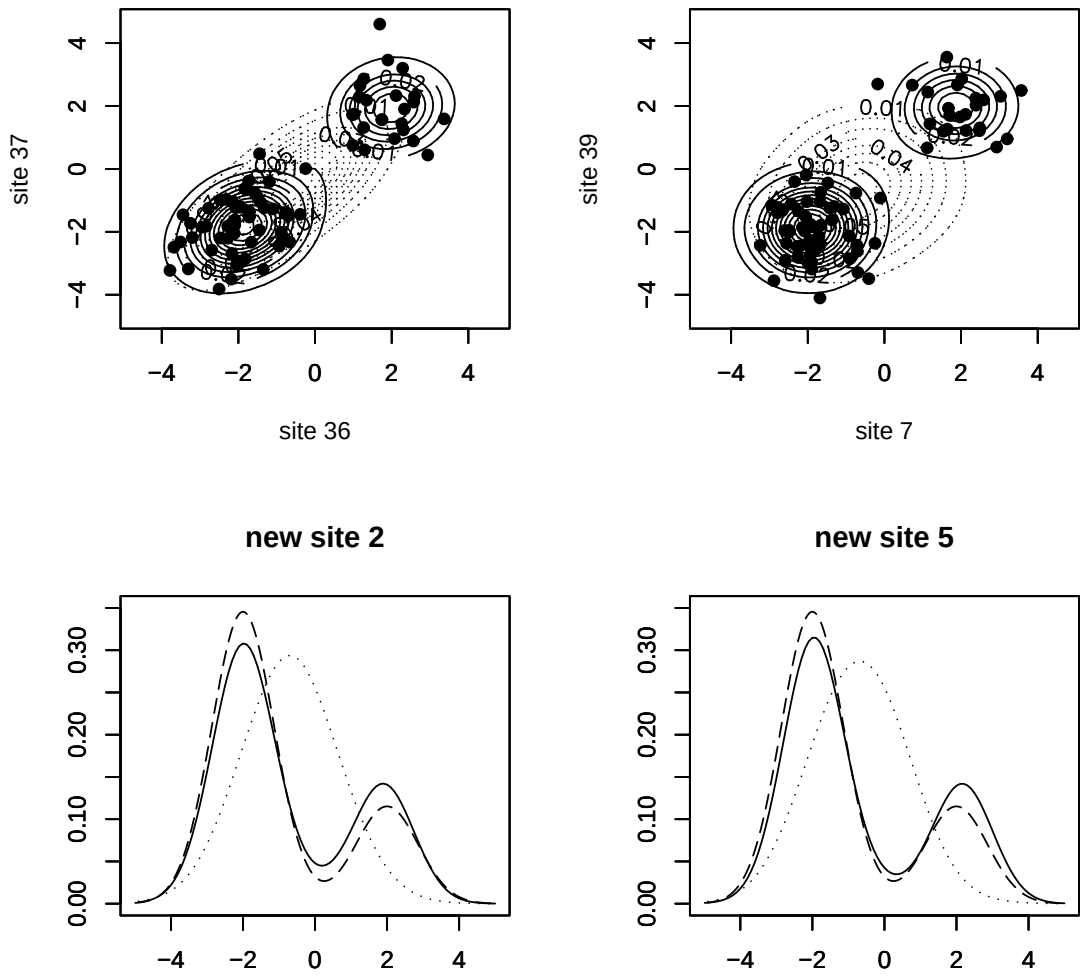


Figure 4.5: Simulated data. Posterior predictive densities under spatial DP mixture model (solid lines) and GP mixture model (dotted lines)

Spatial Dirichlet process models

- *Precipitation data from the Languedoc-Rousillon region in southern France*
- Data were discussed, for example, in Damian, Sampson & Guttorp (2001) $\rightarrow$ original version of the dataset includes 108 altitude-adjusted 10-day aggregated precipitation records for the 39 sites in Figure 4.4
- 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 4.4, 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

Spatial Dirichlet process models

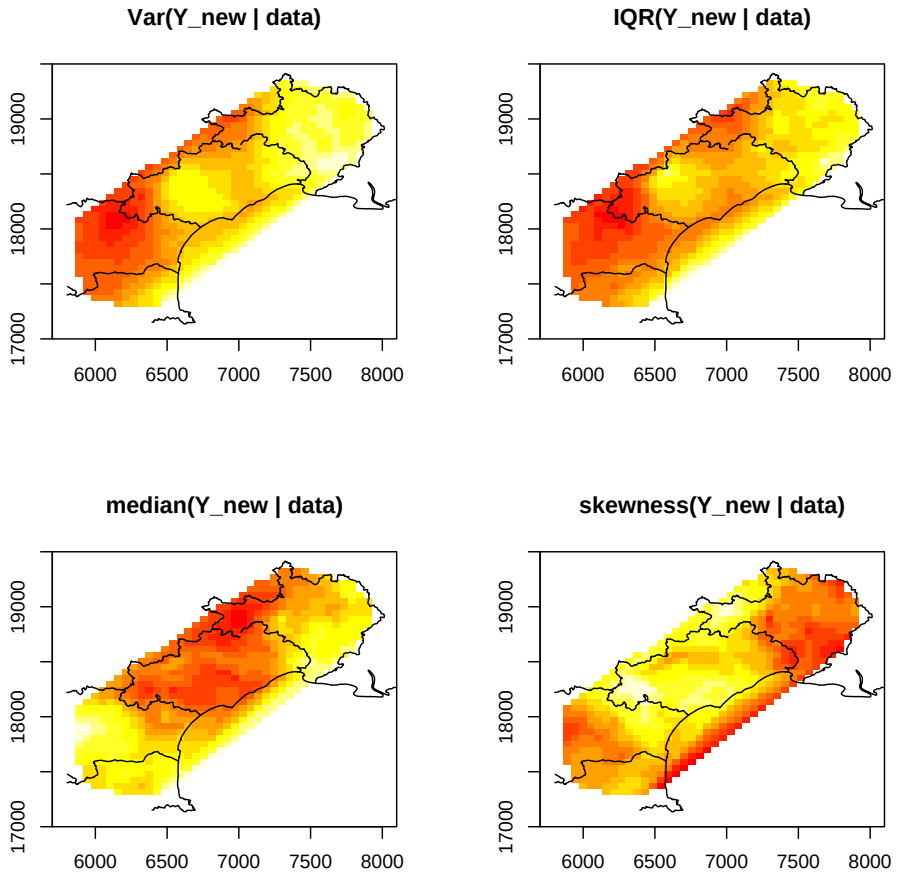


Figure 4.6: French precipitation data. Image plots based on functionals of posterior predictive distributions at observed sites and a number of new sites (darker colors correspond to smaller values)

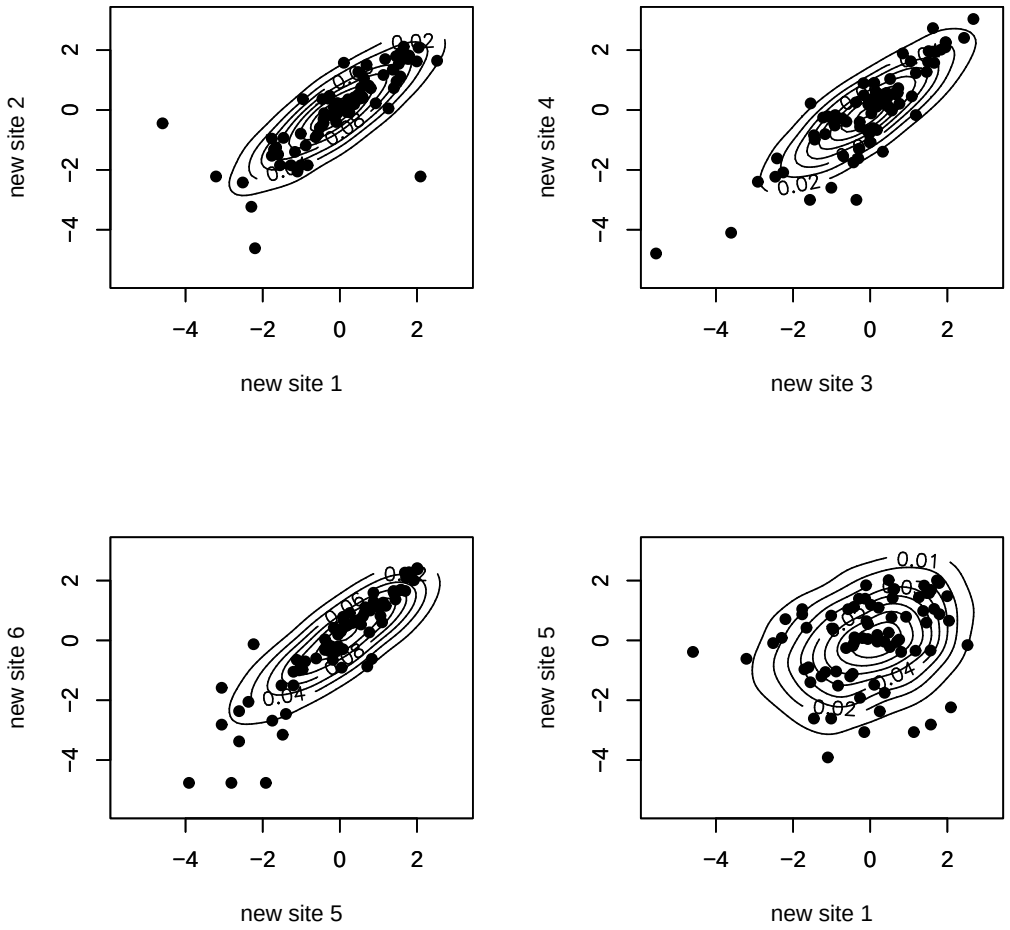


Figure 4.7: French precipitation data. Bivariate posterior predictive densities for pairs of sites (s_4, s_{35}), (s_{29}, s_{30}), (s_{13}, s_{37}) and (s_4, s_{13}) based on model fitted to data after removing sites s_4, s_{35}, s_{29}, s_{30}, s_{13} and s_{37} (overlaid on data observed at the corresponding pairs of sites in the full dataset)

References

- Antoniak, C.E. (1974), "Mixtures of Dirichlet Processes With Applications to Bayesian Nonparametric Problems," *The Annals of Statistics*, 2, 1152-1174.
- Arjas, E., and Gasbarra, D. (1996), "Bayesian inference of survival probabilities, under stochastic ordering constraints," *Journal of the American Statistical Association*, 91, 1101-1109.
- Basu, S., and Mukhopadhyay, S. (2000), "Bayesian analysis of binary regression using symmetric and asymmetric links," *Sankhya, Series B, Indian Journal of Statistics*, 62, 372-387.
- Berry, D.A., and Christensen, R. (1979), "Empirical Bayes Estimation of a Binomial Parameter via Mixtures of a Dirichlet Process," *The Annals of Statistics*, 7, 558-568.
- Bhattacharya, P.K. (1981), "Posterior Distribution of a Dirichlet Process from Quantal Response Data," *The Annals of Statistics*, 9, 803-811.
- Blackwell, D. (1973), "Discreteness of Ferguson Selections," *The Annals of Statistics*, 1, 356-358.
- Blackwell, D., and MacQueen, J.B. (1973), "Ferguson Distributions via Pólya Urn Schemes," *The Annals of Statistics*, 1, 353-355.
- Blei, D.M., and Jordan, M.I. (2006), "Variational inference for Dirichlet process mixtures," *Bayesian Analysis*, 1, 121-144.
- Branscum, A.J., Johnson, W.O., Hanson, T.E., and Gardner, I.A. (2008), "Bayesian semiparametric ROC curve estimation and disease diagnosis," *Statistics in Medicine*, 27, 2474-2496.
- Brunner, L.J. (1992), "Bayesian Nonparametric Methods for Data from a Unimodal Density," *Statistics and Probability Letters*, 14, 195-199.
- Brunner, L.J. (1995), "Bayesian Linear Regression With Error Terms That Have Symmetric Unimodal Densities," *Journal of Nonparametric Statistics*, 4, 335-348.

References

- Brunner, L.J., and Lo, A.Y. (1989), "Bayes Methods for a Symmetric Unimodal Density and its Mode," *The Annals of Statistics*, 17, 1550-1566.

- Burr, D., Doss, H., Cooke, G.E., Goldschmidt-Clermont, P.J. (2003), "A meta-analysis of studies on the association of the platelet PIA polymorphism of glycoprotein IIIa and risk of coronary heart disease," *Statistics in Medicine*, 22, 1741-1760.

- Bush, C.A., and MacEachern, S.N. (1996), "A Semiparametric Bayesian Model for Randomised Block Designs," *Biometrika*, 83, 275-285.

- Cao, G., and West, M. (1996), "Practical Bayesian inference using mixtures of mixtures," *Biometrics*, 52, 1334-1341.

- Carota, C., and Parmigiani, G. (2002), "Semiparametric regression for count data," *Biometrika*, 89, 265-281.

- Chen, Z., and Dunson, D. (2004), "Bayesian estimation of survival functions under stochastic precedence," *Lifetime Data Analysis*, 10, 159-173.

- Chib, S., and Hamilton, B.H. (2002), "Semiparametric Bayes analysis of longitudinal data treatment models," *Journal of Econometrics*, 110, 67-89.

- Cifarelli, D.M., and Regazzini, E. (1978), "Nonparametric statistical problems under partial exchangeability. The use of associative means," (in Italian), *Annali dell' Istituto di Matematica Finanaziaria dell'Universita di Torino, Serie III*, 12, 1-36.

- Connor, R.J., and Mosimann, J.E. (1969), "Concepts of independence for proportions with a generalization of the Dirichlet distribution," *Journal of the American Statistical Association*, 64, 194-206.

- Dahl, D.B. (2005), "Sequentially-Allocated Merge-Split Sampler for Conjugate and Nonconjugate Dirichlet Process Mixture Models," Technical Report.

References

- Damian, D., Sampson, P.D., and Guttorp, P. (2001), "Bayesian Estimation of Semi-parametric Non-stationary Spatial Covariance Structures," *Environmetrics*, 12, 161-178.
- Das, K., and Chattopadhyay, A.K. (2004), "An analysis of clustered categorical data – application in dental health," *Statistics in Medicine*, 23, 2895-2910.
- De Iorio, 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.
- Dey, D., Mueller, P., and Sinha, D. (Editors) (1998). *Practical Nonparametric and Semiparametric Bayesian Statistics*. New York: Springer.
- Diaconis, P., and Ylvisaker, D. (1985), "Quantifying prior opinion," in: J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds. *Bayesian statistics 2* (North-Holland, Amsterdam).
- Disch, D. (1981), "Bayesian Nonparametric Inference for Effective Doses in a Quantal-Response Experiment," *Biometrics*, 37, 713-722.
- Do, K.-A., Müller, P., and Tang, F. (2005), "A Bayesian mixture model for differential gene expression," *Applied Statistics (Journal of the Royal Statistical Society, Series C)*, 54, 627-644.
- Dominici, F., and Parmigiani, G. (2001), "Bayesian Semiparametric Analysis of Developmental Toxicology Data," *Biometrics*, 57, 150-157.
- Duan, J.A., Guindani, M., and Gelfand, A.E. (2007), "Generalized Spatial Dirichlet Process Models," *Biometrika*, 94, 809-825.
- Dunson, D.B. (2005), "Bayesian semiparametric isotonic regression for count data," *Journal of the American Statistical Association*, 100, 618-627.
- Dunson, D.B. and Park, J.-H. (2008), "Kernel stick-breaking processes," *Biometrika*, 95, 307-323.

References

- Dykstra, R.L., and Laud, P. (1981), "A Bayesian Nonparametric Approach to Reliability," *The Annals of Statistics*, 9, 356-367.
- Erkanli, A., Stangl, D. and Müller, P. (1993), "A Bayesian analysis of ordinal data using mixtures," *ASA Proceedings of the Section on Bayesian Statistical Science*, 51-56, American Statistical Association (Alexandria, VA).
- Erkanli, A., Sung, M., Costello, E.J., and Angold, A. (2006), "Bayesian semi-parametric ROC analysis," *Statistics in Medicine*, 25, 3905-3928.
- Escobar, M.D. (1988), "Estimating the Means of Several Normal Populations by Nonparametric Estimation of the Distribution of the Means," Ph.D. dissertation, Yale University.
- Escobar, M.D. (1994), "Estimating Normal Means With a Dirichlet Process Prior," *Journal of the American Statistical Association*, 89, 268-277.
- Escobar, M.D., and West, M. (1995), "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association,* 90, 577-588.
- Fabius, J. (1964), "Asymptotic Behavior of Bayes' Estimates," *The Annals of Mathematical Statistics*, 35, 846-856.
- Ferguson, T.S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, 1, 209-230.
- Ferguson, T.S. (1974), "Prior Distributions on Spaces of Probability Measures," *The Annals of Statistics*, 2, 615-629.
- Ferguson, T.S. (1983), "Bayesian Density Estimation by Mixtures of Normal Distributions," in *Recent Advances in Statistics*, eds. M.H. Rizvi, J.S. Rustagi and D. Siegmund, New York: Academic Press, pp. 287-302.

References

- Ferguson, T.S., and Phadia, E.G. (1979), "Bayesian nonparametric estimation based on censored data," *The Annals of Statistics*, 7, 163-186.
- Freedman, D.A. (1963), "On the Asymptotic Behavior of Bayes' Estimates in the Discrete Case," *The Annals of Mathematical Statistics*, 34, 1386-1403.
- Gasparini, M. (1996), "Bayesian density estimation via Dirichlet density processes," *Journal of Nonparametric Statistics*, 6, 355-366.
- Gelfand, A.E. (1999), "Approaches for Semiparametric Bayesian Regression," in *Asymptotics, Nonparametrics and Time Series*, ed. S. Ghosh, New York: Marcel Dekker, pp. 615-638.
- Gelfand, A.E., and Kuo, L. (1991), "Nonparametric Bayesian Bioassay Including Ordered Polytomous Response," *Biometrika*, 78, 657-666.
- Gelfand, A.E., and Mukhopadhyay, S. (1995), "On Nonparametric Bayesian Inference for the Distribution of a Random Sample," *The Canadian Journal of Statistics*, 23, 411-420.
- Gelfand, A.E., and Kottas, A. (2001), "Nonparametric Bayesian Modeling for Stochastic Order," *Annals of the Institute of Statistical Mathematics*, 53, 865-876.
- Gelfand, A.E., and Kottas, A. (2002), "A Computational Approach for Full Nonparametric Bayesian Inference under Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 11, 289-305.
- Gelfand, A.E., and Kottas, A. (2003), "Bayesian Semiparametric Regression for Median Residual Life," *Scandinavian Journal of Statistics*, 30, 651-665.
- Gelfand, A.E., Kottas, A., and MacEachern, S.N. (2005), "Bayesian Nonparametric Spatial Modeling With Dirichlet Process Mixing," *Journal of the American Statistical Association*, 100, 1021-1035.

References

- Ghosh, J.K., and Ramamoorthi, R.V. (2003). *Bayesian Nonparametrics*. New York: Springer.
- Gopalan, R., and Berry, D.A. (1998), "Bayesian Multiple Comparisons Using Dirichlet Process Priors," *Journal of the American Statistical Association*, 93, 1130-1139.
- Griffin, J.E., and Steel, M.F.J. (2004), "Semiparametric Bayesian inference for stochastic frontier models," *Journal of Econometrics*, 123, 121-152.
- Griffin, J.E., and Steel, M.F.J. (2006), "Order-based dependent Dirichlet processes," *Journal of the American statistical Association*, 101, 179-194.
- Giudici, P., Mezzetti, M., and Muliere, P. (2003), "Mixtures of products of Dirichlet processes for variable selection in survival analysis," *Journal of Statistical Planning and Inference*, 111, 101-115.
- Griffin, J.E., and Steel, M.F.J. (2006), "Order-based dependent Dirichlet processes," *Journal of the American Statistical Association*, 101, 179-194.
- Gutiérrez-Peña, E., and Nieto-Barajas, L.E. (2003), "Bayesian nonparametric inference for mixed Poisson processes," in: J.M. Bernardo, M.J. Bayarri, J.O. Berger, A.P. Dawid, D. Heckerman, A.F.M. Smith and M. West, eds. *Bayesian Statistics 7* (University Press, Oxford), pp. 163-179.
- Guttorp, P. (1995), *Stochastic Modelling of Scientific Data*, Chapman & Hall, Boca Raton.
- Hanson, T. (2006a), "Inference for mixtures of finite Polya trees models," *Journal of the American Statistical Association*, 101, 1548-1565.
- Hanson, T.E. (2006b), "Modeling censored lifetime data using a mixture of gammas baseline," *Bayesian Analysis*, 1, 575-594.
- Hanson, T., and Johnson, W.O. (2002), "Modeling regression error with a mixture of Pólya trees," *Journal of the American Statistical Association*, 97, 1020-1033.

References

- Hanson, T., Branscum, A., and Johnson, W. (2005), "Bayesian Nonparametric Modeling and Data Analysis: An Introduction," in *Bayesian Thinking: Modeling and Computation (Handbook of Statistics, volume 25)*, eds. D.K. Dey and C.R. Rao, Amsterdam: Elsevier, pp. 245-278.

- Hanson, T.E., Kottas, A., and Branscum, A.J. (2008), "Modeling stochastic order in the analysis of receiver operating characteristic data: Bayesian non-parametric approaches," *Journal of the Royal Statistical Society, Series C*, 57, 207-225.

- Hasegawa, H., and Kozumi, H. (2003), "Estimation of Lorenz curves: a Bayesian nonparametric approach," *Journal of Econometrics*, 115, 277-291.

- Hirano, K. (2002), "Semiparametric Bayesian inference in autoregressive panel data models," *Econometrica*, 70, 781-799.

- Hjort, N.L. (1996), "Bayesian Approaches to Non- and Semiparametric Density Estimation," in *Bayesian Statistics 5, Proceedings of the Fifth Valencia International Meeting*, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, Oxford: Oxford Clarendon Press, pp. 223-253.

- Hjort, N.L. (2000), "Bayesian Analysis for a Generalised Dirichlet Process Prior," Statistical Research Report, Department of Mathematics, University of Oslo.

- Hoff, P. (2003), "Bayesian methods for partial stochastic orderings," *Biometrika*, 90, 303-317.

- Hoff, P. (2005), "Subset clustering of binary sequences, with an application to genomic abnormality data," *Biometrics*, 61, 1027-1036.

- Ibrahim, J.G., Chen, M-H., and Sinha, D. (2001), *Bayesian survival analysis*, New York: Springer.

- Ishwaran, H., and James, L.F. (2001), "Gibbs Sampling Methods for Stick-Breaking Priors," *Journal of the American Statistical Association*, 96, 161-173.

- Ishwaran, H., and James, L.F. (2002), "Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information," *Journal of Computational and Graphical Statistics*, 11, 1-26.

References

- Ishwaran, H., and James, L.F. (2004), "Computational methods for multiplicative intensity models using weighted gamma processes: Proportional hazards, marked point processes, and panel count data," *Journal of the American Statistical Association*, 99, 175-190.
- Ishwaran, H., and Zarepour, M. (2000), "Markov Chain Monte Carlo in Approximate Dirichlet and Beta Two-parameter Process Hierarchical Models," *Biometrika*, 87, 371-390.
- Jain, S., and Neal, R. M. (2004), "A Split-Merge Markov Chain Monte Carlo Procedure for the Dirichlet Process Mixture Model," *Journal of Computational and Graphical Statistics*, 13, 158-182.
- Jain, S., and Neal, R. M. (2007), "Splitting and Merging Components of a Nonconjugate Dirichlet Process Mixture Model," (with discussion), *Bayesian Analysis*, 2, 445-500.
- Jasra, A., Holmes, C.C., and Stephens, D.A. (2005), "Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling," *Statistical Science*, 20, 50-67.
- Johnson, V.E., and Albert, J.H. (1999), *Ordinal Data Modeling*, New York: Springer.
- Kalbfleisch, J.D. (1978), "Non-parametric Bayesian Analysis of Survival Time Data," *Journal of the Royal Statistical Society, Ser. B*, 40, 214-221.
- Kingman, J.F.C. (1993), *Poisson Processes*, Clarendon Press, Oxford.
- Kleinman, K.P., and Ibrahim, J.G. (1998), "A semi-parametric Bayesian approach to generalized linear mixed models," *Statistics in Medicine*, 17, 2579-2596.
- Kottas, A. (2006a), "Dirichlet process mixtures of Beta distributions, with applications to density and intensity estimation," *Proceedings of the Workshop on Learning with Nonparametric Bayesian Methods*, 23rd International Conference on Machine Learning, Pittsburgh, PA.
- Kottas, A. (2006b), "Nonparametric Bayesian Survival Analysis using Mixtures of Weibull Distributions," *Journal of Statistical Planning and Inference*, 136, 578-596.

References

- Kottas, A. (2009), "Bayesian semiparametric inference under stochastic precedence order constraints, with applications in epidemiology and survival analysis," Technical Report, Department of Applied Mathematics and Statistics, University of California, Santa Cruz.

- Kottas, A., and Behseta, S. (2009), "Bayesian nonparametric modeling for comparison of single-neuron firing intensities," to appear in *Biometrics*.

- Kottas, A., and Gelfand, A.E. (2001a), "Modeling Variability Order: A Semiparametric Bayesian Approach," *Methodology and Computing in Applied Probability*, 3, 427-442.

- Kottas, A., and Gelfand, A.E. (2001b), "Bayesian Semiparametric Median Regression Modeling," *Journal of the American Statistical Association*, 96, 1458-1468.

- Kottas, A., and Krnjajić, M. (2009), "Bayesian Semiparametric Modeling in Quantile Regression," *Scandinavian Journal of Statistics*, 36, 297-319.

- Kottas, A., and Sansó, B. (2007), "Bayesian mixture modeling for spatial Poisson process intensities, with applications to extreme value analysis," *Journal of Statistical Planning and Inference*, 137, 3151-3163.

- Kottas, A., Branco, M.D., and Gelfand, A.E. (2002), "A Nonparametric Bayesian Modeling Approach for Cytogenetic Dosimetry," *Biometrics*, 58, 593-600.

- Kottas, A., Duan, J.A., and Gelfand, A.E. (2008), "Modeling Disease Incidence Data with Spatial and Spatio-temporal Dirichlet Process Mixtures," *Biometrical Journal*, 50, 29-42.

- Kottas, A., Müller, P., and Quintana, F. (2005), "Nonparametric Bayesian Modeling for Multivariate Ordinal Data," *Journal of Computattional and Graphical Statistics*, 14, 610-625.

- Krnjajić, M., Kottas, A., and Draper, D. (2008), "Parametric and Nonparametric Bayesian Model Specification: A Case Study Involving Models for Count Data," *Computational Statistics & Data Analysis*, 52, 2110-2128.

References

- Kuo, L. (1983), "Bayesian Bioassay Design," *The Annals of Statistics*, 11, 886-895.
- Kuo, L. (1986a), "A Note on Bayes Empirical Bayes Estimation by Means of Dirichlet Processes," *Statistics and Probability Letters*, 4, 145-150.
- Kuo, L. (1986b), "Computations of Mixtures of Dirichlet Processes," *SIAM Journal on Scientific and Statistical Computing*, 7, 60-71.
- Kuo, L. (1988), "Linear Bayes Estimators of the Potency Curve in Bioassay," *Biometrika*, 75, 91-96.
- Kuo, L., and Ghosh, S.K. (1997). Bayesian nonparametric inference for nonhomogeneous Poisson processes. Technical Report, Department of Statistics, University of Connecticut.
- Kuo, L., and Mallick, B. (1997), "Bayesian Semiparametric Inference for the Accelerated Failure-Time Model," *The Canadian Journal of Statistics*, 25, 457-472.
- Kuo, L., and Smith, A.F.M. (1992), "Bayesian Computations in Survival Models via the Gibbs Sampler," in *Survival Analysis: State of the Art*, eds. J.P. Klein and P.K. Goel, Dodrecht: Kluwer Academic Publishers, pp. 11-24.
- Lavine, M. (1992), "Some Aspects of Polya Tree Distributions for Statistical Modelling," *The Annals of Statistics*, 20, 1222-1235.
- Lavine, M. (1994), "More Aspects of Polya Tree Distributions for Statistical Modelling," *The Annals of Statistics*, 22, 1161-1176.
- Lavine, M., and Mockus, A. (1995), "A Nonparametric Bayes Method for Isotonic Regression," *Journal of Statistical Planning and Inference*, 46, 235-248.
- Lau, J.W., and Green, P.J. (2007), "Bayesian Model-Based Clustering Procedures," *Journal of Computational and Graphical Statistics*, 16, 526-558.
- Lawless, J.F. (1982), *Statistical models and methods for lifetime data*, New York: Wiley.

References

- Lee, J., and Berger, J.O. (1999), "Semiparametric Bayesian Analysis of Selection Models," *Journal of the American Statistical Association*, 96, 1397-1409.
- Leslie, D.S., Kohn, R., and Nott, D.J. (2007), "A general approach to heteroscedastic linear regression," *Statistics and Computing*, 17, 131-146.
- Lijoi, A., Mena, R.H., and Prünster, I. (2007), "Controlling the reinforcement in Bayesian non-parametric hierarchical models," *Journal of the Royal Statistical Society, Series B*, 69, 715-740.
- Liu, J.S. (1996), "Nonparametric Hierarchical Bayes via Sequential Imputations," *The Annals of Statistics*, 24, 911-930.
- Lo, A.Y. (1984), "On a Class of Bayesian Nonparametric Estimates: I. Density Estimates," *The Annals of Statistics*, 12, 351-357.
- Lo, A.Y., and Weng, C.-S. (1989), "On a class of Bayesian nonparametric estimates: II. Hazard rate estimates," *Annals of the Institute of Statistical Mathematics*, 41, 227-245.
- MacEachern, S.N. (1999), "Dependent Nonparametric Processes," in *ASA Proceedings of the Section on Bayesian Statistical Science*, Alexandria, VA: American Statistical Association, pp. 50-55.
- MacEachern, S.N. (2000), "Dependent Dirichlet Processes," Technical Report, Department of Statistics, The Ohio State University.
- MacEachern, S.N., and Müller, P. (1998), "Estimating Mixture of Dirichlet Process Models," *Journal of Computational and Graphical Statistics,* 7, 223-238.
- MacEachern, S.N., Clyde, M., and Liu, J.S. (1999), "Sequential Importance Sampling for Nonparametric Bayes Models: The Next Generation," *The Canadian Journal of Statistics*, 27, 251-267.
- 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.

References

- Mallick, B.K., and Walker, S.G. (1997), "Combining Information from Several Experiments With Nonparametric Priors," *Biometrika*, 84, 697-706.
- Mauldin, R.D., Sudderth, W.D., and Williams, S.C. (1992), "Polya trees and random distributions," *The Annals of Statistics*, 20, 1203-1221.
- Merrick, J.R.W., Soyer, R., and Mazzuchi, T.A. (2003), "A Bayesian semiparametric analysis of the reliability and maintenance of machine tools," *Technometrics*, 45, 58-69.
- Mira, A., and Petrone, S. (1996), "Bayesian Hierarchical Nonparametric Inference for Change-Point Problems," in *Bayesian Statistics 5, Proceedings of the Fifth Valencia International Meeting*, eds. J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith, Oxford: Oxford Clarendon Press, pp. 693-703.
- Moller, J., and Waagepetersen, R.P. (2004), *Statistical Inference and Simulation for Spatial Point Processes*, Chapman & Hall, Boca Raton.
- Mukhopadhyay, S. (2000), "Bayesian Nonparametric Inference on the Dose Level With Specified Response Rate," *Biometrics*, 56, 220-226.
- Mukhopadhyay, S., and Gelfand, A.E. (1997), "Dirichlet Process Mixed Generalized Linear Models," *Journal of the American Statistical Association*, 92, 633-639.
- Muliere, P., and Tardella, L. (1998), "Approximating Distributions of Random Functionals of Ferguson-Dirichlet Priors," *The Canadian Journal of Statistics*, 26, 283-297.
- Muliere, P., and Walker, S. (1997a), "A Bayesian Non-parametric Approach to Survival Analysis Using Polya Trees," *Scandinavian Journal of Statistics*, 24, 331-340.
- Muliere, P., and Walker, S. (1997b), "A Bayesian nonparametric approach to determining a maximum tolerated dose," *Journal of Statistical Planning and Inference*, 61, 339-353.
- Müller, P., and Quintana, F.A. (2004), "Nonparametric Bayesian Data Analysis," *Statistical Science*, 19, 95-110.

References

- Müller, P., and Roeder, K. (1997), "A Bayesian Semiparametric Model for Case-control Studies With Errors in Variables," *Biometrika*, 84, 523-537.
- Müller, P., and Rosner, G.L. (1997), "A Bayesian Population Model With Hierarchical Mixture Priors Applied to Blood Count Data," *Journal of the American Statistical Association*, 92, 1279-1292.
- Müller, P., Erkanli, A., and West, M. (1996), "Bayesian Curve Fitting Using Multivariate Normal Mixtures," *Biometrika*, 83, 67-79.
- Müller, P., Quintana, F.A., and Rosner, G. (2004), "A method for combining inference across related nonparametric Bayesian models," *Journal of the Royal Statistical Society, Ser. B*, 66, 735-749.
- Müller, P., West, M., and MacEachern, S. (1997), "Bayesian Models for Non-linear Autoregressions," *Journal of Time Series Analysis*, 18, 593-614.
- Müller, P., Rosner, G.L., De Iorio, M., and MacEachern, S. (2005), "A nonparametric Bayesian model for inference in related longitudinal studies," *Applied Statistics (Journal of the Royal Statistical Society, Series C)*, 54, 611-626.
- Neal, R.M. (2000), "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 9, 249-265.
- Newton, M.A., and Zhang, Y. (1999), "A Recursive Algorithm for Nonparametric Analysis With Missing Data," *Biometrika*, 86, 15-26.
- Ongaro, A., and Cattaneo, C. (2004), "Discrete random probability measures: a general framework for nonparametric Bayesian inference," *Statistics and Probability Letters*, 67, 33-45.
- Paddock, S. (2002), "Bayesian nonparametric multiple imputation of partially observed data with ignorable nonresponse," *Biometrika*, 89, 529-538.
- Paddock, S.M., Ruggeri, F., Lavine, M., and West, M. (2003), "Randomized Polya tree models for nonparametric Bayesian inference," *Statistica Sinica*, 13, 443-460.

References

- Papaspiliopoulos, O., and Roberts, G.O. (2008), "Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models," *Biometrika*, 95, 169-186.
- Pitman, J., and Yor, M. (1997), "The Two-Parameter Poisson-Dirichlet Distribution Derived from a Stable Subordinator," *The Annals of Probability*, 25, 855-900.
- Quintana, F.A. (1998), "Nonparametric Bayesian Analysis for Assessing Homogeneity in $k \times l$ Contingency Tables With Fixed Right Margin Totals," *Journal of the American Statistical Association*, 93, 1140-1149.
- Richardson, S., and Green, P.J. (1997), "On Bayesian Analysis of Mixtures With an Unknown Number of Components" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 59, 731-792.
- Rodriguez, A., Dunson, D.B., and Gelfand, A.E. (2008), "The nested Dirichlet process," (with discussion), *Journal of the American Statistical Association*, 103, 1131-1154.
- Rodriguez, A., Dunson, D.B., and Gelfand, A.E. (2009), "Bayesian nonparametric functional data analysis through density estimation," *Biometrika*, 96, 149-162.
- Sethuraman, J., and Tiwari, R.C. (1982), "Convergence of Dirichlet Measures and the Interpretation of their Parameter," in *Statistical Decision Theory and Related Topics III*, eds. S. Gupta and J.O. Berger, New York: Springer-Verlag, 2, pp. 305-315.
- Sethuraman, J. (1994), "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, 4, 639-650.
- Shaked, M. and Shanthikumar, J.G. (1994), *Stochastic Orders and Their Applications*, Boston: Academic Press.

References

- Sinha, D., and Dey, D.K. (1997), "Semiparametric Bayesian Analysis of Survival Data," *Journal of the American Statistical Association*, 92, 1195-1212.

- Stephens, M. (2000), "Bayesian Analysis of Mixture Models With an Unknown Number of Components - An Alternative to Reversible Jump Methods," *The Annals of Statistics*, 28, 40-74.

- Taddy, M., and Kottas, A. (2009a), "A Bayesian Nonparametric Approach to Inference for Quantile Regression," to appear in *Journal of Business and Economic Statistics*.

- Taddy, M., and Kottas, A. (2009b), "Dirichlet Process Mixture Modeling for Marked Poisson Processes," Technical Report UCSC-SOE-09-31, Department of Applied Mathematics and Statistics, University of California, Santa Cruz.

- Teh, Y.W., Jordan, M.I., Beal, M.J., and Blei, D.M. (2006), "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, 101, 1566-1581.

- Tomlinson, G., and Escobar, M. (1999), "Analysis of Densities," Research Report, Department of Public Health Sciences, University of Toronto.

- Walker, S.G., and Mallick, B.K. (1999), "A Bayesian semiparametric accelerated failure time model," *Biometrics*, 55, 477-483.

- Walker, S.G., Damien, P., Laud, P.W., and Smith, A.F.M. (1999), "Bayesian Nonparametric Inference for Random Distributions and Related Functions" (with discussion), *Journal of the Royal Statistical Society, Ser. B*, 61, 485-527.

- West, M., Müller, P., and Escobar, M.D. (1994), "Hierarchical Priors and Mixture Models, With Application in Regression and Density Estimation," in *Aspects of Uncertainty: A Tribute to D.V. Lindley*, eds. A.F.M. Smith and P. Freeman, New York: Wiley, pp. 363-386.

- Wolpert, R.L., and Ickstadt, K. (1998), "Poisson/Gamma random field models for spatial statistics," *Biometrika*, 85, 251-267.