

Instructor: Athanasios Kottas, Department of Applied Mathematics and Statistics, University of California, Santa Cruz

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

Outline

1. Bayesian nonparametrics
2. The Dirichlet process
3. Dose-response modeling with Dirichlet process priors
4. Bayesian nonparametric modeling for cytogenetic dosimetry
5. Semiparametric regression for categorical responses
6. Other Bayesian nonparametric approaches

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 $\{h_k(x) : k = 1, \dots, K\}$ 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}

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

- Main focus: Modeling **random distributions**
- Distributions can be over scalars, vectors, even over a stochastic process
- Much more than cdfs
- 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

- General review papers on Bayesian nonparametrics: Walker, Damien, Laud and Smith (1999); Müller and Quintana (2004); Hanson, Branscum and Johnson (2005)
- Review papers on specific application areas of Bayesian nonparametric and semiparametric methods: Hjort (1996); Sinha and Dey (1997); Gelfand (1999)
- Books: Dey, Müller and Sinha (1998) (edited volume with a collection of papers, mainly, on applications of Bayesian nonparametrics); Ghosh and Ramamoorthi (2003) (emphasis on theoretical development of Bayesian nonparametric priors)

The Dirichlet process

- The DP is characterized by two parameters:
 - Q_0 a specified probability measure on $(\mathcal{X}, \mathcal{B})$ (equivalently, G_0 a specified distribution function on \mathcal{X})
 - 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, \dots, B_k of \mathcal{X} ,

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

→ 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
 → also, the $B_i, i = 1, \dots, k$, define a measurable partition if $B_i \in \mathcal{B}$, they are pairwise disjoint, and their union is \mathcal{X}

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

Recall the definition of the Dirichlet distribution

- Start with independent rvs $Z_j \sim \text{gamma}(a_j, 1), j = 1, \dots, k$ (with $a_j > 0$)
- Define $Y_j = Z_j / (\sum_{\ell=1}^k Z_\ell)$, for $j = 1, \dots, k$
- Then $(Y_1, \dots, Y_k) \sim \text{Dirichlet}(a_1, \dots, a_k)$ (distribution singular w.r.t. Lebesgue measure on R^k , since $\sum_{j=1}^k Y_j = 1$)
- (Y_1, \dots, 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: $E(Y_j) = a_j / \sum_{\ell=1}^k a_\ell, E(Y_j^2) = a_j(a_j + 1) / \{\sum_{\ell=1}^k a_\ell(1 + \sum_{\ell=1}^k a_\ell)\}$, and, for $i \neq j, E(Y_i Y_j) = a_i a_j / \{\sum_{\ell=1}^k a_\ell(1 + \sum_{\ell=1}^k a_\ell)\}$
- Note that for $k = 2, \text{Dirichlet}(a_1, a_2) \equiv \text{Beta}(a_1, a_2)$

- 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

$$E(Q(B)) = Q_0(B)$$

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 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 $x_1 < x_2 < \dots < x_k$ in \mathcal{X}
- Then, the random vector $(G(x_1), G(x_2) - G(x_1), \dots, G(x_k) - G(x_{k-1}), 1 - G(x_k))$ follows a Dirichlet distribution with parameters $(\alpha G_0(x_1), \alpha(G_0(x_2) - G_0(x_1)), \dots, \alpha(G_0(x_k) - G_0(x_{k-1})), \alpha(1 - G_0(x_k)))$
- Hence, if (u_1, u_2, \dots, u_k) is a draw from this Dirichlet distribution, then $(u_1, \dots, \sum_{j=1}^i u_j, \dots, \sum_{j=1}^k u_j)$ is a draw from the distribution of $(G(x_1), \dots, G(x_i), \dots, G(x_k))$
- Example (Figure 1): $\mathcal{X} = (0, 1)$, $G_0(x) = x$, $x \in (0, 1)$ (Unif(0, 1) base distribution)

- 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

$$E(G(x)) = G_0(x)$$

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

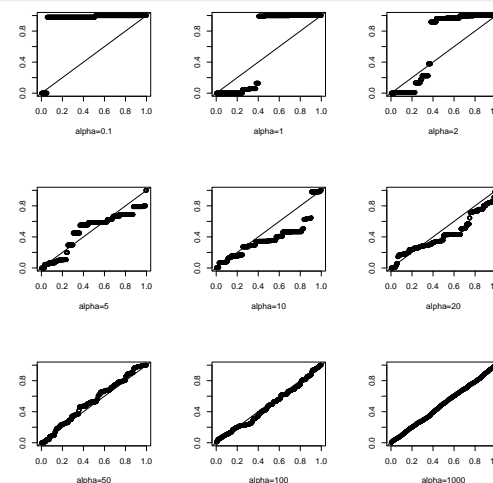


Figure 1: Cdf sample paths from a $\text{DP}(\alpha, G_0 = \text{Unif}(0, 1))$ prior, for different values of α . The solid line denotes the cdf of G_0 .

• **Constructive definition of the DP**

(Sethuraman and Tiwari, 1982; Sethuraman, 1994)

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

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

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

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

→ 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

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

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

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

The Dirichlet process

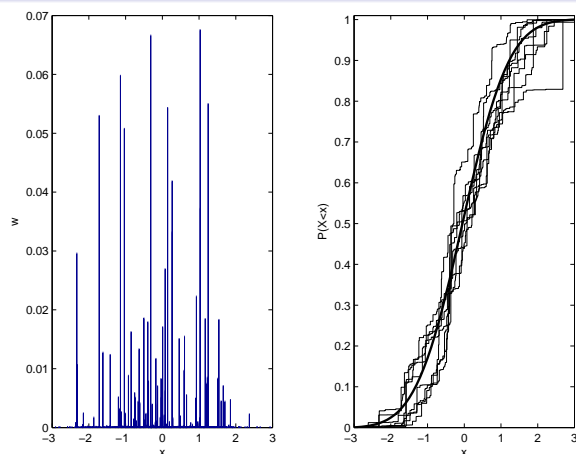


Figure 2: Illustration for a DP with $G_0 = N(0, 1)$ and $\alpha = 20$. In the left panel, the spiked lines are located at 1000 sampled values of x drawn from $N(0, 1)$ with heights given by the weights, ω_ℓ , calculated using the stick-breaking algorithm (a truncated version so that the weights sum 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.

The Dirichlet process

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

- **Pólya urn characterization of the DP**

(Blackwell and MacQueen, 1973)

→ if, for $i = 1, \dots, n$, $x_i | G$ 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

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

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

* $x_1 \sim G_0$, and

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

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

$$E(G(t) | y_1, \dots, y_n) = \frac{\alpha}{\alpha + n} G_0(t) + \frac{n}{\alpha + n} G_n(t)$$

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)

→ for small α relative to n , little weight is placed on the prior guess G_0

→ for large α relative to n , little weight is placed on the data

→ 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)

→ note that, $\lim_{\alpha \rightarrow 0} E(G(t) | y_1, \dots, y_n) = G_n(t)$

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

(Ferguson, 1973)

→ let G denote the random distribution function for the following results

→ if the observations $y_i | G$ are i.i.d. from G , $i = 1, \dots, n$, 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

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 and Zarepour, 2000) is defined by choosing $z_r \sim \text{Beta}(a, b)$
- If $z_r \sim \text{Beta}(1 - a, b + ra)$, $r = 1, 2, \dots$, for some $a \in [0, 1)$ and $b \in (-a, \infty)$, we obtain the two-parameter Poisson-Dirichlet process (e.g., Pitman and Yor, 1997)
- The general case, $z_r \sim \text{Beta}(a_r, b_r)$ (Ishwaran and 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

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

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 and 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)

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
 - x_i : dose levels (with $x_1 < x_2 < \dots < x_k$)
 - n_i : number of subjects at dose level i
 - 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 $\mathcal{X} \subseteq \mathcal{R}$

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 $p_i = F(x_i)$, $i = 1, \dots, k$
- 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, \dots, p_k) is an ordered Dirichlet distribution, i.e., $(p_1, p_2 - p_1, \dots, p_k - p_{k-1}, 1 - p_k)$ follows a Dirichlet distribution with parameters $(\alpha F_0(x_1), \alpha(F_0(x_2) - F_0(x_1)), \dots, \alpha(F_0(x_k) - F_0(x_{k-1})), \alpha(1 - F_0(x_k)))$
- The posterior for F is a **mixture of Dirichlet processes** (Antoniak, 1974)
 - a random distribution H follows a mixture of DPs, denoted by $H \sim \int \text{DP}(\alpha, H_0(\phi)) P(d\phi)$, if the random variable ϕ has distribution P and then conditional on ϕ , H follows a DP with base distribution $H_0(\phi)$
 - posterior distribution is difficult to work with analytically (Antoniak obtained point estimate when $k = 2$) — Markov chain Monte Carlo (MCMC) techniques enable full inference (as in, e.g., Gelfand and Kuo, 1991; Mukhopadhyay, 2000)

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

- Bayesian nonparametric modeling for polytomous response (Kottas, Branco and 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 $\mathbf{q}_\ell = (q_{\ell,1}, \dots, q_{\ell,k})$, $\ell = 1, 2$, where $q_{\ell,i} = G_\ell(x_i)$

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

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

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

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 $p(\mathbf{q}_1, \mathbf{q}_2 \mid \text{data})$
- 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 , $i = 1, \dots, k$
- 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

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

where $p(q_{\ell,0} \mid \mathbf{q}_\ell)$ 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 $\mathbf{y}_0 = (y_{01}, y_{02}, y_{03})$ — 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

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

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

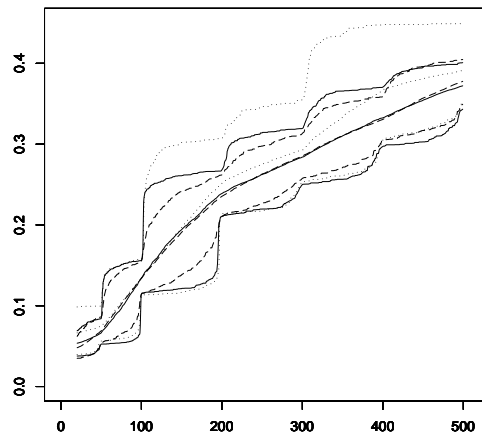


Figure 3: For the data in Table 1, point and 95% pointwise interval posterior estimates for the probability of at least one MN vs dose under $\alpha_1 = \alpha_2 = 0.1$ (dotted lines), 1 (solid lines) and 10 (dashed lines).

Data Illustrations

Table 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., $\hat{\eta}_{i1} = y_{i1}/(y_{i1} + y_{i2} + y_{i3})$, and at least one micronuclei, i.e., $\hat{\eta}_{i2} = (y_{i1} + y_{i2})/(y_{i1} + y_{i2} + y_{i3})$.

i	Dose (cGy)	y_{i1}	y_{i2}	y_{i3}	$\hat{\eta}_{i1}$	$\hat{\eta}_{i2}$
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

- 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

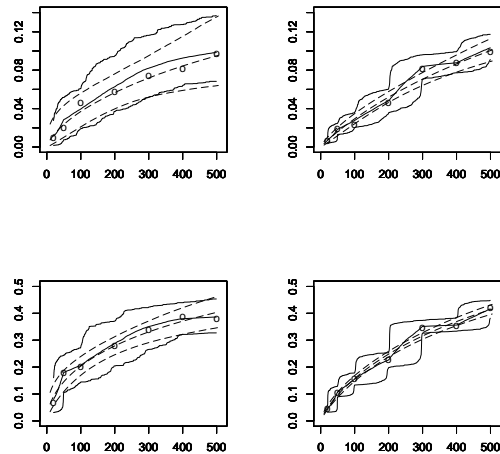


Figure 4: Simulation 1. Posterior inference for F_1 (upper panels) and F_2 (lower panels) under the parametric (dashed lines) and nonparametric (solid lines) model. “o” denotes the observed data. The left and right panels correspond to the data set with the smaller and large sample sizes, respectively.

Semiparametric regression for categorical responses

5. Semiparametric regression for categorical responses

- Application of DP-based modeling to semiparametric regression with categorical responses
- Categorical responses y_i , $i = 1, \dots, N$ (e.g., counts or proportions)
- Covariate vector \mathbf{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 ($k = 1, \dots, K$) 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 , $k = 1, \dots, K$

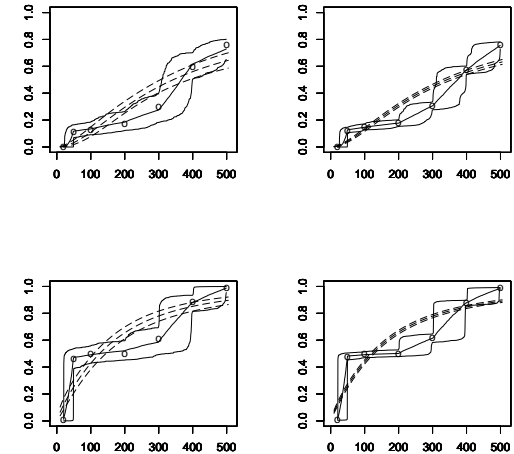


Figure 5: Simulation 2. Posterior inference for F_1 (upper panels) and F_2 (lower panels) under the parametric (dashed lines) and nonparametric (solid lines) model. “o” denotes the observed data. The left and right panels correspond to the data set with the smaller and large sample sizes, respectively.

Semiparametric regression for categorical responses

- *Product of mixtures of Dirichlet processes prior* (Cifarelli and Regazzini, 1978) for the cell-specific random distributions F_k , $k = 1, \dots, K$
 - conditionally on hyperparameters α_k and θ_k , the F_k are assigned independent $\text{DP}(\alpha_k, F_{0k}(\cdot; \theta_k))$ priors, where, in general, $\theta_k = (\theta_{1k}, \dots, \theta_{Dk})$
 - the F_k are related by modeling the α_k ($k = 1, \dots, K$) and/or the θ_{dk} ($k = 1, \dots, K; d = 1, \dots, D$) as linear combinations of the predictors (through specified link functions h_d , $d = 0, 1, \dots, D$)
 - $h_0(\alpha_k) = \mathbf{x}_k^T \boldsymbol{\gamma}$, $k = 1, \dots, K$
 - $h_d(\theta_{dk}) = \mathbf{x}_k^T \boldsymbol{\beta}_d$, $k = 1, \dots, K; d = 1, \dots, D$
 - (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 $\{F_1, \dots, F_K\}$, though a weaker type of dependence than more recent approaches building on dependent DP priors (MacEachern, 2000)

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

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

$$y_i | \{F_1, \dots, F_K\} \sim \prod_{i=1}^N F_{k(i)}(y_i)$$

$$F_k | \alpha_k, \theta_k \stackrel{\text{ind.}}{\sim} \text{DP}(\alpha_k, \text{Poisson}(\cdot; \theta_k)), \quad k = 1, \dots, K$$

$$\log(\alpha_k) = \mathbf{x}_k^T \boldsymbol{\gamma} \quad \log(\theta_k) = \mathbf{x}_k^T \boldsymbol{\beta}, \quad k = 1, \dots, K$$

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

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

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 $E(G(B_0)) = 0.5 = G_0(B_0)$
- 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)
 - 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, 2006) or Pólya trees with “jittered” partitions (Paddock et al., 2003)

6. Other Bayesian nonparametric approaches

Pólya tree priors

- Pólya tree processes (Ferguson, 1974; Mauldin, Sudderth and 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:
 - $G(B_0) = \Pr(\theta \in B_0) = \nu_0$, where $\nu_0 \sim \text{Beta}(\alpha_0, \alpha_1)$
 - $\Pr(\theta \in B_{00} | \theta \in B_0) = \nu_{00}$, with $\nu_{00} \sim \text{Beta}(\alpha_{00}, \alpha_{01})$, so $G(B_{00}) = \nu_0 \nu_{00}$
 - 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

- 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
- Modeling applications with Pólya tree priors: survival analysis (Muliere and Walker, 1997a; Walker and Mallick, 1999); bioassay modeling (Muliere and Walker, 1997b); median regression (Hanson and Johnson, 2002); multiple imputation with partially observed data (Paddock, 2002); ROC data analysis (Branscum et al., 2008; Hanson, Kottas and 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 and Phadia, 1979) (i.e., $Z(\cdot)$ has independent increments, and is, almost surely, non-decreasing, right continuous, with $Z(0) = 0$ and $\lim_{t \rightarrow \infty} Z(t) = \infty$; 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)
 - Consider an arbitrary finite partition of R^+ , $0 = a_0 < a_1 < a_2 \dots < a_k < a_{k+1} = \infty$
 - Let $q_l = \Pr(T \in [a_{l-1}, a_l] | T \geq a_{l-1})$ and let $r_l = -\log(1 - q_l)$
 - then $\sum_{l=1}^k r_l = -\log \Pr(T \geq a_k) = Z(a_k)$

References

References

- Antoniak, C.E. (1974), "Mixtures of Dirichlet Processes With Applications to Bayesian Nonparametric Problems," *The Annals of Statistics*, 2, 1152-1174.
- 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.
- 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.
- Carota, C., and Parmigiani, G. (2002), "Semiparametric regression for count data," *Biometrika*, 89, 265-281.
- 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 Finanziaria dell'Universita di Torino, Serie III*, 12, 1-36.
- 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.
- Disch, D. (1981), "Bayesian Nonparametric Inference for Effective Doses in a Quantal-Response Experiment," *Biometrics*, 37, 713-722.

- 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 and Laud, 1981)

References

- 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.
- Dykstra, R.L., and Laud, P. (1981), "A Bayesian Nonparametric Approach to Reliability," *The Annals of Statistics*, 9, 356-367.
- 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., 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.
- 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., Kottas, A., and MacEachern, S.N. (2005), "Bayesian Nonparametric Spatial Modeling With Dirichlet Process Mixing," *Journal of the American Statistical Association*, 100, 1021-1035.

- Ghosh, J.K., and Ramamoorthi, R.V. (2003). *Bayesian Nonparametrics*. New York: Springer.
- 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.
- Hanson, T. (2006), "Inference for mixtures of finite Polya trees models," *Journal of the American Statistical Association*, 101, 1548-1565.
- 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.
- 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.
- 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.
- 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 Zarepour, M. (2000), "Markov Chain Monte Carlo in Approximate Dirichlet and Beta Two-parameter Process Hierarchical Models," *Biometrika*, 87, 371-390.
- Kalbfleisch, J.D. (1978), "Non-parametric Bayesian Analysis of Survival Time Data," *Journal of the Royal Statistical Society, Ser. B*, 40, 214-221.
- 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.
- Kuo, L. (1983), "Bayesian Bioassay Design," *The Annals of Statistics*, 11, 886-895.
- Kuo, L. (1988), "Linear Bayes Estimators of the Potency Curve in Bioassay," *Biometrika*, 75, 91-96.
- 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.
- 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.
- 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.

- Mauldin, R.D., Sudderth, W.D., and Williams, S.C. (1992), "Polya trees and random distributions," *The Annals of Statistics*, 20, 1203-1221.
- 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.
- Mukhopadhyay, S. (2000), "Bayesian Nonparametric Inference on the Dose Level With Specified Response Rate," *Biometrics*, 56, 220-226.
- 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.
- 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.

- Pitman, J., and Yor, M. (1997), "The Two-Parameter Poisson-Dirichlet Distribution Derived from a Stable Subordinator," *The Annals of Probability*, 25, 855-900.
- 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.
- 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.
- Sinha, D., and Dey, D.K. (1997), "Semiparametric Bayesian Analysis of Survival Data," *Journal of the American Statistical Association*, 92, 1195-1212.
- 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.