AMS 241: Bayesian Nonparametric Methods
Notes 1 – Dirichlet process priors

Instructor: Athanasios Kottas

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

Outline

1. Bayesian nonparametrics: introduction and motivation
2. Definition of the Dirichlet process
3. Constructive definition of the Dirichlet process
4. Pólya urn representation of the Dirichlet process
5. Posterior inference
6. Mixtures of Dirichlet processes
7. Applications

Parametric vs. nonparametric Bayes: A simple example

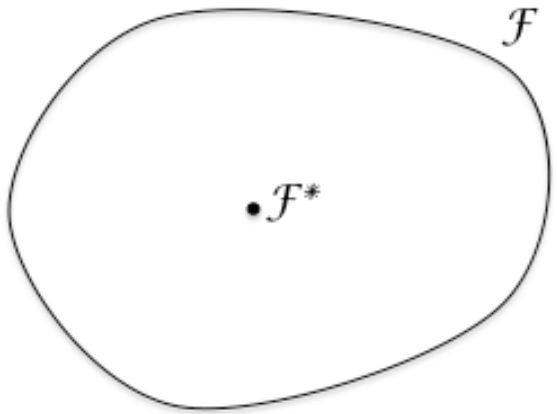
- Let $y_i \in \mathcal{Y}$, $y_i \mid F \sim_{iid} F$, $F \in \mathcal{F}^*$,

$$\mathcal{F}^* = \{\mathrm{N}(y \mid \mu, \tau^2), \mu \in \mathbb{R}, \tau \in \mathbb{R}^+\}.$$

- In this **parametric** specification a prior on $\mathcal{F}^*$ boils down to a prior on (μ, τ^2).
- However, $\mathcal{F}^*$ is tiny compared to

$$\mathcal{F} = \{\text{All distributions on } \mathcal{Y}\}.$$

- **Nonparametric Bayes** involves priors on much larger subsets of $\mathcal{F}$ (infinite-dimensional spaces).
- One handy way to do this is to use stochastic processes.



Bayesian nonparametrics

- Priors on spaces of functions, $\{g(\cdot) : g \in \mathcal{G}\}$ (infinite-dimensional spaces) vs usual parametric priors on Θ, where $g(\cdot) \equiv g(\cdot; \theta)$, $\theta \in \Theta$
- In certain applications, we may seek more structure, e.g., monotone regression functions or unimodal error densities.
- Even though we focus on priors for distributions (priors for density or distribution functions), the methods are more widely useful: hazard or cumulative hazard function, intensity functions, link function, calibration function ...
- More generally, enriching usual parametric models, typically leading to semiparametric models.
- Wandering nonparametrically near a standard class.
- *Bayesian nonparametrics*, an oxymoron? very different from classical nonparametric estimation techniques.

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., $g(\cdot)$ may be a realization from a Gaussian process over $\mathcal{X}$
- Main focus: Modeling **random distributions**
- Distributions can be over scalars, vectors, even over a stochastic process (much more than c.d.f.s).

Bayesian nonparametrics

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

Some references

- General review papers on Bayesian nonparametrics: Walker, Damien, Laud and Smith (1999); Müller and Quintana (2004); Hanson, Branscum and Johnson (2005); Müller and Mitra (2013).
- Review papers on specific application areas of Bayesian nonparametric and semiparametric methods: Hjort (1996); Sinha and Dey (1997); Gelfand (1999).
- Books and edited volumes: Dey, Müller and Sinha (1998); Ghosh and Ramamoorthi (2003); Hjort, Holmes, Müller and Walker (2010); Müller and Rodriguez (2013); Müller, Quintana, Jara and Hanson (2015).
- Software: functions for nonparametric Bayesian inference are spread over various `R` packages. Only comprehensive package we are aware of is the `DPpackage`.

The Dirichlet process as a model for random distributions

- 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., $\mathcal{X} = \mathbb{R}$, or $\mathbb{R}^+$, or $\mathbb{R}^d$), equipped with a σ-field $\mathcal{B}$ of subsets of $\mathcal{X}$ (e.g., the Borel σ-field for $\mathcal{X} \subseteq \mathbb{R}^d$)
- 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 distributions.
- 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 $\mathcal{X} \subseteq \mathbb{R}^d$, equivalently, random c.d.f.s on $\mathcal{X}$

Constructing priors for spaces of distributions

- Defining the prior as a stochastic process with sample paths that are distributions on $(\mathcal{X}, \mathcal{B})$.

- Consistency conditions for the finite dimensional distributions (f.d.d.s) (see Ferguson, 1973, and Walker et al., 1999).

- Let $\mathcal{G}_Q$ be the space of probability measures (distributions) Q on $(\mathcal{X}, \mathcal{B})$. Consider a system of f.d.d.s for $(Q(B_{1,1}), ..., Q(B_{m,k}))$ for each finite collection $B_{1,1}, ..., B_{m,k}$ of pairwise disjoint sets in $\mathcal{B}$. If:
  - $Q(B)$ is a random variable taking values in $[0, 1]$, for all $B \in \mathcal{B}$;
  - $Q(\mathcal{X}) = 1$ almost surely; and
  - $(Q(\cup_{i=1}^{k} B_{1,i}), ..., Q(\cup_{i=1}^{k} B_{m,i}))$ and $(\sum_{i=1}^{k} Q(B_{1,i}), ..., \sum_{i=1}^{k} Q(B_{m,i}))$ are equal in distribution

  then, there exists a unique (random) probability measure on $\mathcal{G}_Q$ with these f.d.d.s.

Motivating the construction of 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 the next two slides).
- With the Dirichlet process we further generalize to infinite-dimensional spaces.

Properties of the Dirichlet distribution

- Start with independent random variables

$$Z_j \sim \text{gamma}(a_j, 1), \qquad j = 1, ..., k,$$

  with $a_j > 0$.
- Define

$$Y_j = \frac{Z_j}{\sum_{\ell=1}^{k} Z_\ell}, \qquad j = 1, ..., k.$$

- Then $(Y_1, ..., Y_k) \sim \text{Dirichlet}(a_1, ..., a_k)$.

- This distribution is singular w.r.t. Lebesgue measure on $\mathbb{R}^k$, since $\sum_{j=1}^{k} Y_j = 1$.

Properties of the Dirichlet distribution

- $(Y_1, ..., Y_{k-1})$ has density

$$\frac{\Gamma\left(\sum_{j=1}^{k} a_j\right)}{\prod_{j=1}^{k} \Gamma(a_j)} \left(1 - \sum_{j=1}^{k-1} y_j\right)^{a_k - 1} \prod_{j=1}^{k-1} y_j^{a_j - 1}.$$

- Note that for $k = 2$, Dirichlet$(a_1, a_2) \equiv$ Beta(a_1, a_2).
- The moments of the Dirichlet distribution are:

$$\mathrm{E}(Y_j) = \frac{a_j}{\sum_{\ell=1}^{k} a_\ell}, \qquad \mathrm{E}(Y_j^2) = \frac{a_j(a_j+1)}{\sum_{\ell=1}^{k} a_\ell (1 + \sum_{\ell=1}^{k} a_\ell)},$$

$$\mathrm{E}(Y_i Y_j) = \frac{a_i a_j}{\sum_{\ell=1}^{k} a_\ell (1 + \sum_{\ell=1}^{k} a_\ell)}, \quad \text{for } i \neq j.$$

- We can think about the Dirichlet as having *two parameters*:
  - $g = \{a_j / (\sum_{\ell=1}^{k} a_\ell) : j = 1, ..., k\}$, the mean vector.
  - $\alpha = \sum_{\ell=1}^{k} a_\ell$, a concentration parameter controlling its variance.

Definition of the Dirichlet process

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

- **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 $\mathcal{X}$,

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

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

Definition of the Dirichlet process

- Regarding existence of the DP as a random probability measure, the key property of the Dirichlet distribution is "additivity", which results from the additive property of the gamma distribution: for independent r.v.s $Z_r \sim \text{gamma}(a_r, 1)$, for $r = 1, 2$, $Z_1 + Z_2 \sim \text{gamma}(a_1 + a_2, 1)$.

- Additive property of the Dirichlet distribution:
  if $(Y_1, ..., Y_k) \sim \text{Dirichlet}(a_1, ..., a_k)$, and $m_1, ..., m_M$ are integers such that $0 < m_1 < ... < m_M = k$, then the random vector

$$(\sum_{i=1}^{m_1} Y_i, \sum_{i=m_1+1}^{m_2} Y_i, ..., \sum_{i=m_{M-1}+1}^{m_M} Y_i)$$

  has a Dirichlet$(\sum_{i=1}^{m_1} a_i, \sum_{i=m_1+1}^{m_2} a_i, ..., \sum_{i=m_{M-1}+1}^{m_M} a_i)$ distribution.

- Using the additivity property of the Dirichlet distribution, the Kolmogorov consistency conditions can be established for the f.d.d.s of $(Q(B_1), ..., Q(B_k))$ in the DP definition (refer to Lemma 1 in Ferguson, 1973).

Interpreting the parameters of 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

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

- Q_0 plays the role of the *center* of the DP (also referred to as baseline probability measure, or baseline 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).

Interpreting the parameters of the Dirichlet process

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

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

and thus

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

- **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.

Simulating c.d.f. realizations from a Dirichlet process

- The definition can be used to simulate sample paths (which are distribution functions) from the DP – this is convenient when $\mathcal{X} \subseteq \mathbb{R}$.
- Consider any grid of points $x_1 < x_2 < ... < x_k$ in $\mathcal{X} \subseteq \mathbb{R}$.
- 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 parameter vector

$$(\alpha G_0(x_1), \alpha(G_0(x_2) - G_0(x_1)), ..., \alpha(G_0(x_k) - G_0(x_{k-1})), \alpha(1 - G_0(x_k)))$$

- 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$(0, 1)$ centering distribution).

Simulating c.d.f. realizations from a Dirichlet process

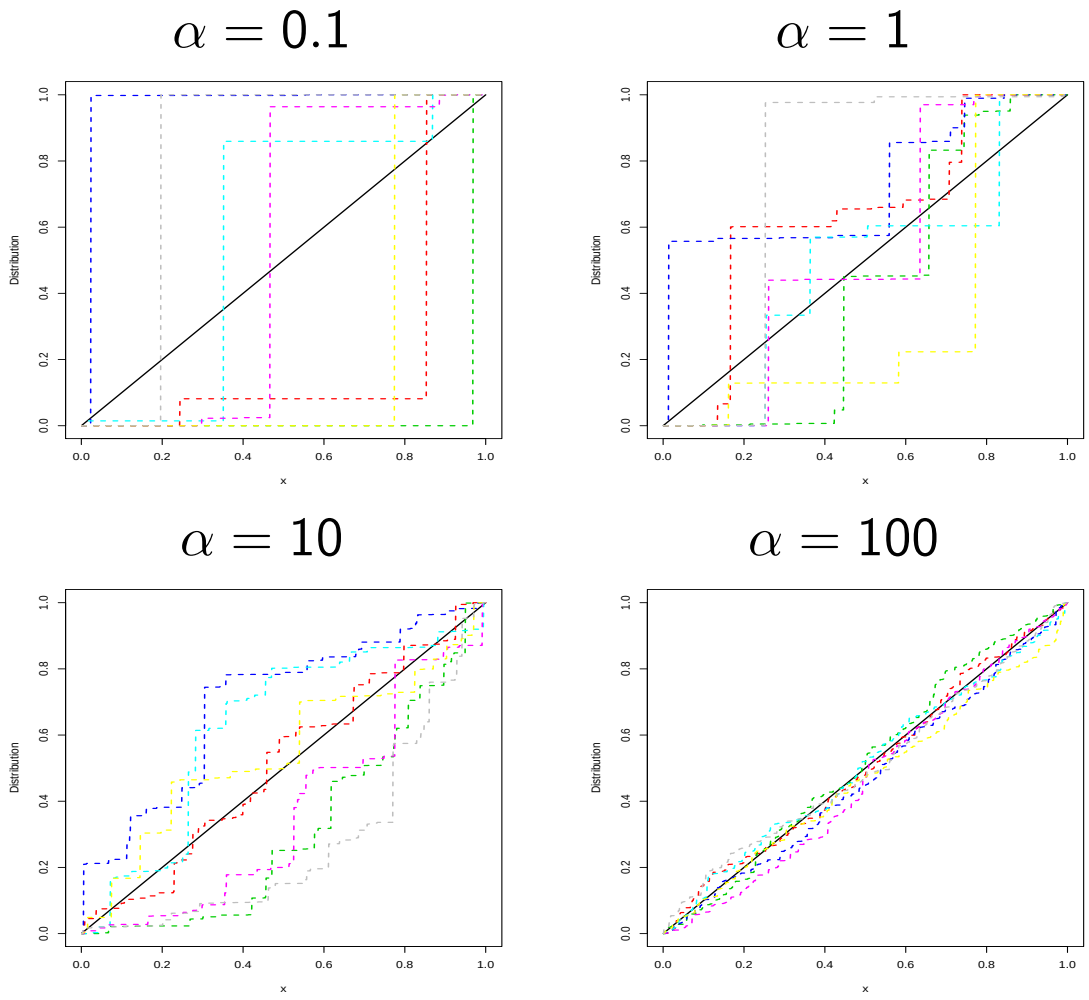


Figure 1.1: C.d.f. realizations from a DP(α, G_0 = Unif(0, 1)) for different α values. The solid black line corresponds to the baseline uniform c.d.f., while the dashed colored lines represent multiple realizations.

Further references on the Dirichlet process

- Early work on study of theoretical properties of the DP; e.g., Korwar and Hollander (1973), James and Mosimann (1980), Hannum, Hollander and Langberg (1981), Doss and Sellke (1982), Lo (1983).

- The mean functional, $\mu(G) = \int t \mathrm{d}G(t)$, $G \sim \mathrm{DP}(\alpha, G_0)$, has received special attention.
  - It can be shown that if G_0 has finite mean, then $\mu(G)$ is (almost surely) finite. In this case, $\mathrm{E}(\mu(G)) = \mu(G_0) = \int t \mathrm{d}G_0(t)$.
  - The distribution of $\mu(G)$ has been studied by Yamato (1984), Cifarelli and Regazzini (1990), Diaconis and Kemperman (1996), and Regazzini, Guglielmi and Di Nunno (2002).

- An extensive review of the work on the DP up to 1990 can be found in Ferguson, Phadia and Tiwari (1992).

Constructive definition of the DP

- Due to Sethuraman and Tiwari (1982) and Sethuraman (1994).
- Let $\{z_r : r = 1, 2, ...\}$ and $\{\vartheta_\ell : \ell = 1, 2, ...\}$ be independent sequences of i.i.d. random variables
  - $z_r \sim \text{Beta}(1, \alpha)$, $r = 1, 2, ...$
  - $\vartheta_\ell \sim G_0$, $\ell = 1, 2, ...$
- Define $\omega_1 = z_1$ and $\omega_\ell = z_\ell \prod_{r=1}^{\ell-1}(1 - z_r)$, for $\ell = 2, 3, ...$
- Then, a realization G from $\text{DP}(\alpha, G_0)$ is (almost surely) of the form

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

where $\delta_a(\cdot)$ denotes a point mass at a.

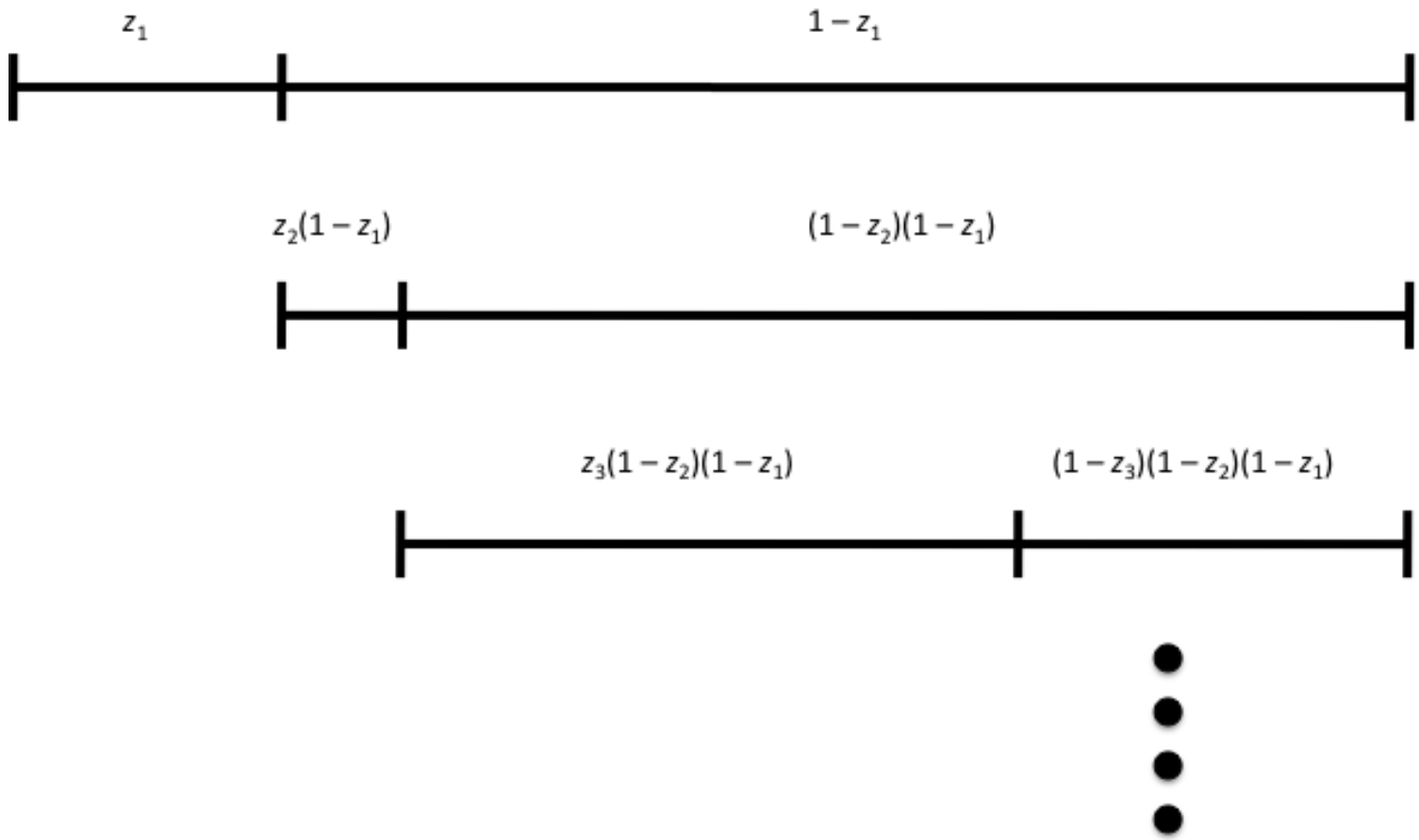
Constructive definition of the DP

- 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.
  - The associated weights ω_ℓ are defined using the *stick-breaking* construction.
- This is not as restrictive as it might sound: Any distribution on $\mathbb{R}^d$ can be approximated arbitrarily well using a countable mixture of point masses.
- The realizations we showed before already hinted at this fact.
- 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 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 ....
- It can be shown that $\sum_{\ell=1}^{\infty} \omega_\ell = 1$ (almost surely).

The stick-breaking construction



More on the constructive definition of the DP

- 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 c.d.f. sample paths.
- For example, a possible approximation is $G_J = \sum_{j=1}^{J} p_j \delta_{\vartheta_j}$, with $p_j = \omega_j$ for $j = 1, ..., J-1$, and $p_J = 1 - \sum_{j=1}^{J-1} \omega_j = \prod_{r=1}^{J-1}(1 - z_r)$.
- To specify J, a simple approach involves working with the expectation for the partial sum of the stick-breaking weights:

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

Hence, J could be chosen such that $\{\alpha/(\alpha+1)\}^J = \varepsilon$, for small ε.

More on the constructive definition of the DP

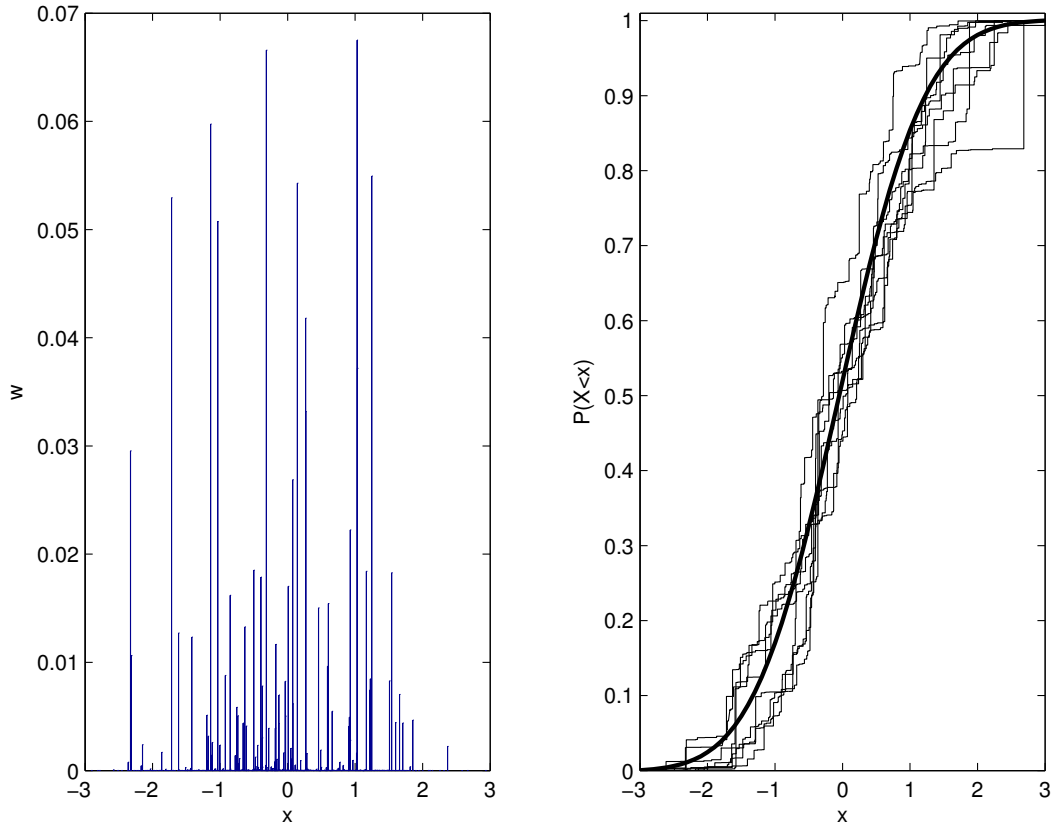


Figure 1.2: Illustration for a DP with $G_0 = \text{N}(0, 1)$ and $\alpha = 20$. In the left panel, the spiked lines are located at 1000 N(0, 1) draws with heights given by the (truncated) stick-breaking weights. These spikes are then summed to generate one c.d.f. sample path. The right panel shows 8 such sample paths indicated by the lighter jagged lines. The heavy smooth line indicates the N(0, 1) c.d.f.

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 $z_r \sim \text{beta}(1 - a, b + ra)$, for $r = 1, 2, \ldots$ and 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).
- The probit stick-breaking process, where $z_r = \Phi(x_r)$ with $x_r \sim N(\mu, \sigma^2)$ and Φ denoting the standard normal c.d.f. (Rodríguez and Dunson, 2011).

Further extensions based on the DP constructive definition

The constructive definition of the DP has motivated several of its extensions, including:

- 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

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

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

  (Blackwell and MacQueen, 1973).

- 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.

Pólya urn characterization of the DP

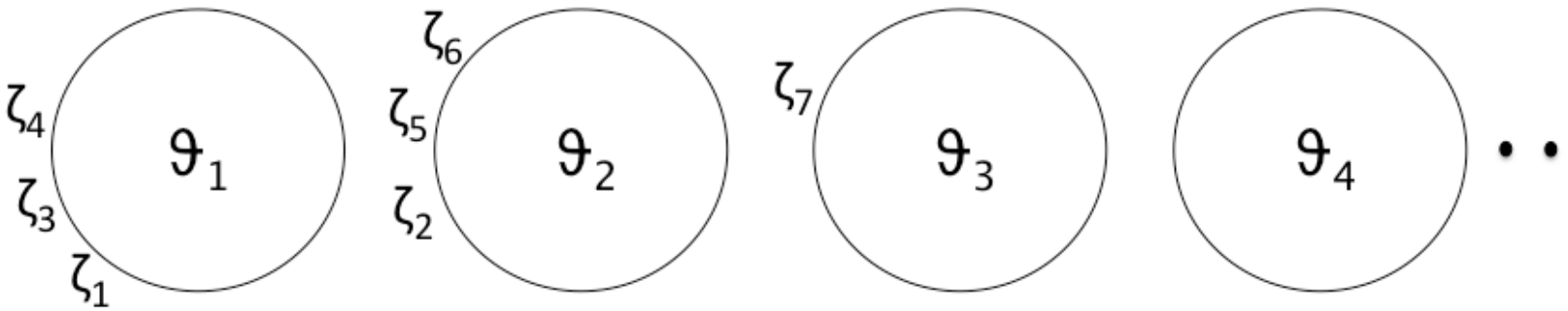
- The forward direction described above (i.e., starting with the DP prior for G) is readily established using results from Ferguson (1973).
- Blackwell and MacQueen (1973) proved the other direction, thus, characterizing the DP as the de Finetti measure for *Pólya sequences*.
- A sequence of r.v.s, $\{X_n : n \geq 1\}$, (w.l.o.g. on $\mathbb{R}$) is a Pólya sequence with parameters G_0 (a distribution on $\mathbb{R}$) and α (a positive scalar parameter) if for any measurable $B \subset \mathbb{R}$, $\Pr(X_1 \in B) = G_0(B)$, and $\Pr(X_{n+1} \in B \mid X_1, ..., X_n) = (\alpha + n)^{-1}\{\alpha G_0(B) + \sum_{i=1}^{n} \delta_{X_i}(B)\}$ (where $\delta_{X_i}(B) = 1$ if $X_i \in B$, and $\delta_{X_i}(B) = 0$ otherwise).
- If $\{X_n : n \geq 1\}$ is a Pólya sequence with parameters α and G_0, then:
  - $(\alpha + n)^{-1}\{\alpha G_0 + \sum_{i=1}^{n} \delta_{X_i}\}$ converges almost surely (as $n \to \infty$) to a discrete distribution G
  - $G \sim \text{DP}(\alpha, G_0)$
  - $X_1, X_2, ... \mid G$ are independently distributed according to G.

The Chinese restaurant process

The Pólya urn characterization of the DP can be visualized using the Chinese restaurant analogy:

- A customer arriving at the restaurant joins a table that already has some customers, with probability proportional to the number of people in the table, or takes the first seat at a new table with probability proportional to α.

- All customers sitting in the same table share a dish.



Prior to posterior updating with DP priors

- In what follows, G denotes the random distribution function.
- Ferguson (1973) has shown that 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.

Prior to posterior updating with DP priors

- For example, the posterior mean 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).

  - 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.
  - Hence, 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).
  - However, taking α very small in order to be "noninformative" is *very* dangerous; recall that α controls both the variance and the extent of discreteness for the DP prior.

C.d.f. estimation using DP priors

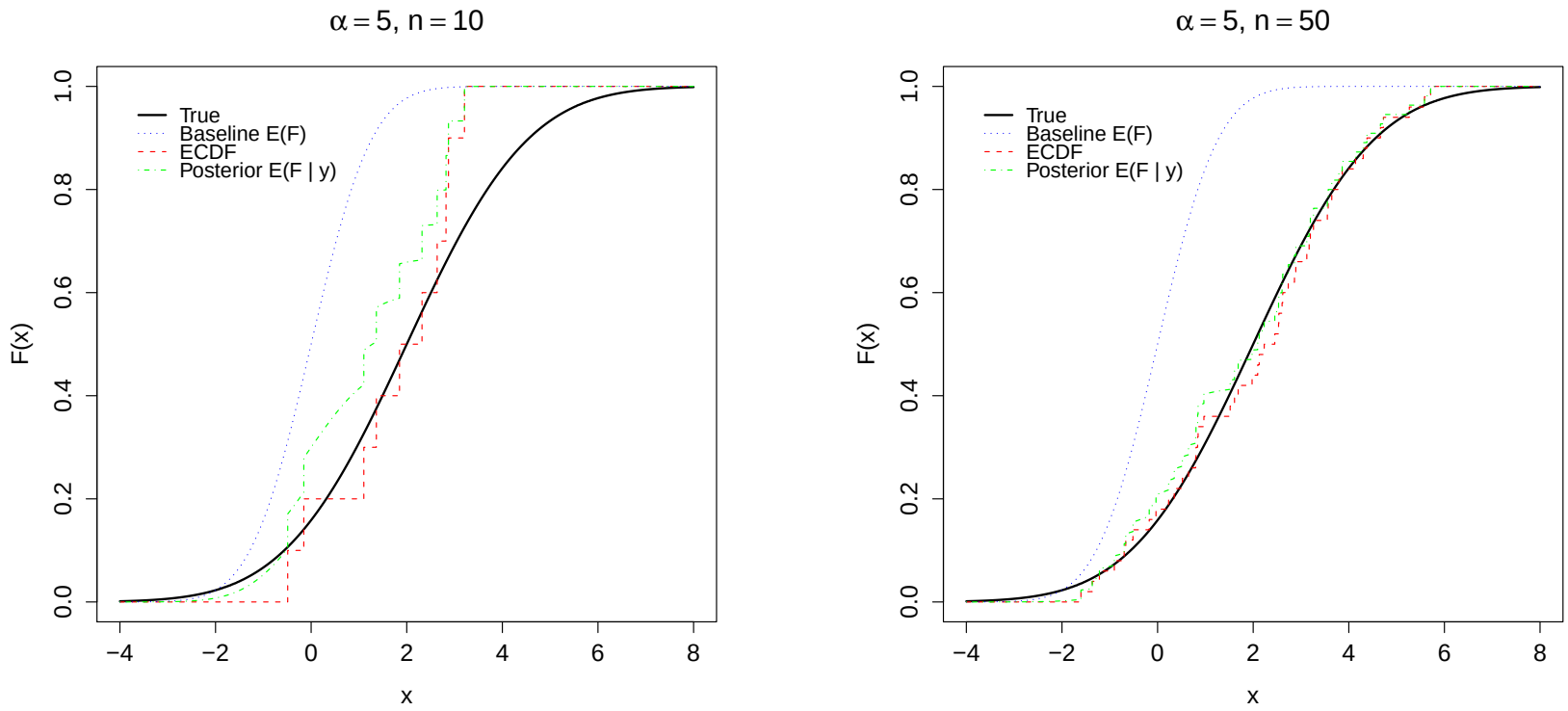


Figure 1.3: Estimating the distribution function under a DP prior, using simulated data. Both the true distribution generating the data and the baseline distribution are Gaussian. The left panel corresponds to a sample of $n = 10$ observations while the right panel corresponds to a sample of $n = 50$ observations.

Some of the early references on inference under DP priors

- Construction of confidence bands for the c.d.f. and interval estimates for the associated mean and quantiles (Breth, 1978, 1979).
- Inference for the survival function based on right censored data (Susarla and van Ryzin, 1976, 1978; Blum and Susarla, 1977) and on grouped data (Johnson and Christensen, 1986).
- Semiparametric survival regression through the accelerated failure time model (Christensen and Johnson, 1988; Johnson and Christensen, 1989). Inference scope extended through posterior simulation (Kuo and Smith, 1992).
- Variants of the DP can be found in Doss (1985a,b) and Newton, Czado and Chappell (1996), including applications to median estimation and binary regression, respectively.

Mixtures of Dirichlet processes

- A random distribution G follows a mixture of Dirichlet processes (MDP) (Antoniak, 1974) if it arises from a DP, but now conditionally on random DP prior parameters (random α and/or G_0).
- The MDP structure extends the DP to a hierarchical setting:

$$G \mid \alpha, \boldsymbol{\psi} \sim \text{DP}(\alpha, G_0(\cdot \mid \boldsymbol{\psi})),$$

  where (parametric) priors are added to the precision parameter α and/or the parameters of the centering distribution, $\boldsymbol{\psi}$.

- Mixtures of Dirichlet processes are different from Dirichlet process mixture models, $f(\cdot \mid G) = \int k(\cdot \mid \boldsymbol{\theta}) \, \mathrm{d}G(\boldsymbol{\theta})$, where k is a parametric kernel density, and $G \sim \text{DP}(\alpha, G_0)$.
  - However, there are important connections: the posterior distribution for G follows the MDP structure (Antoniak, 1974).

Inference for discrete distributions using MDP priors

- The MDP can be used as a prior model for discrete distributions F.
- As an example, consider a discrete distribution with support on $\{0, 1, 2, ...\}$, with observed count responses, data $= \{y_i : i = 1, ..., n\}$.
- MDP prior model with Poisson centering distribution:

$$\begin{aligned} y_i \mid F &\overset{i.i.d.}{\sim} F, \quad i = 1, ..., n \\ F \mid \alpha, \lambda &\sim \text{DP}(\alpha, F_0(\cdot) = \text{Poisson}(\cdot \mid \lambda)) \\ \alpha, \lambda &\sim \pi(\alpha)\pi(\lambda) \end{aligned}$$

- Using results from Antoniak (1974), the joint posterior distribution for F and (α, λ) can be developed through a DP for the conditional posterior of F given (α, λ), and the marginal posterior for (α, λ).
  - Hence, the marginal posterior distribution for F follows the MDP structure, and thus, the MDP is also a conjugate prior.

Inference for discrete distributions using MDP priors

- Joint posterior: $p(F, \alpha, \lambda \mid \text{data}) = p(\alpha, \lambda \mid \text{data}) p(F \mid \alpha, \lambda, \text{data})$
  $\propto \pi(\alpha)\pi(\lambda) L(\alpha, \lambda; \text{data}) p(F \mid \alpha, \lambda, \text{data})$
- Conditional posterior: $p(F \mid \alpha, \lambda, \text{data}) = \text{DP}(\alpha + n, \tilde{F}_0)$, where

$$\tilde{F}_0(y) = \frac{\alpha}{\alpha + n} F_0(y \mid \lambda) + \frac{1}{\alpha + n} \sum_{i=1}^{n} 1_{[y_i, \infty)}(y)$$

- Marginal likelihood (expression specific to DP priors with discrete F_0):

$$L(\alpha, \lambda; \text{data}) \propto \frac{\alpha^{n^*}}{\alpha^{(n)}} \prod_{j=1}^{n^*} f_0(y_j^* \mid \lambda) \{\alpha f_0(y_j^* \mid \lambda) + 1\}^{(n_j - 1)}$$

  - $f_0(\cdot \mid \lambda)$ is the p.m.f. of $F_0(\cdot \mid \lambda)$
  - n^* is the number of distinct values in $(y_1, ..., y_n)$
  - $\{y_j^* : j = 1, ..., n^*\}$ are the distinct values in $(y_1, ..., y_n)$
  - $n_j = |\{i : y_i = y_j^*\}|$, for $j = 1, ..., n^*$
  - notation: $z^{(m)} = z(z+1) \times ... \times (z+m-1)$, for $m > 0$, with $z^{(0)} = 1$

Inference for discrete distributions using MDP priors

- Posterior simulation from $p(F, \alpha, \lambda \mid \text{data})$ through:
  - MCMC sampling from $p(\alpha, \lambda \mid \text{data}) \propto \pi(\alpha)\pi(\lambda)L(\alpha, \lambda; \text{data})$; and
  - simulation from $p(F \mid \alpha, \lambda, \text{data})$, using any of the DP definitions.

- Posterior predictive distribution:

$$\Pr(Y = y \mid \text{data}) = \mathrm{E}\{\Pr(Y = y \mid F) \mid \text{data}\}, \quad y = 0, 1, 2, \ldots$$

  - for $y \geq 1$, $\mathrm{E}\{\Pr(Y = y \mid F) \mid \text{data}\} = \mathrm{E}\{F(y) - F(y-1) \mid \text{data}\}$
  - $\mathrm{E}\{\Pr(Y = 0 \mid F) \mid \text{data}\} = \mathrm{E}\{F(1) - \Pr(Y = 1 \mid F) \mid \text{data}\} = \mathrm{E}\{F(1) \mid \text{data}\} - \Pr(Y = 1 \mid \text{data})$

- For any y, the posterior distribution for the random c.d.f. at y, $F(y)$, can be sampled using the DP definition:

$$p(F(y) \mid \text{data}) = \iint p(F(y) \mid \alpha, \lambda, \text{data}) p(\alpha, \lambda \mid \text{data}) \, \mathrm{d}\alpha \mathrm{d}\lambda$$

where $p(F(y) \mid \alpha, \lambda, \text{data})$ is a Beta distribution with parameters $(\alpha + n)\tilde{F}_0(y)$ and $(\alpha + n)(1 - \tilde{F}_0(y))$.

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 $\mathbf{x}_i$ 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 ($k = 1, ..., K$) 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 and Regazzini, 1978) for the cell-specific random distributions F_k, $k = 1, ..., K$:
  - 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, $\boldsymbol{\theta}_k = (\theta_{1k}, ..., \theta_{Dk})$
  - 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, $d = 0, 1, ..., D$)
  - $h_0(\alpha_k) = \mathbf{x}_k^T \boldsymbol{\gamma}$, $k = 1, ..., K$
  - $h_d(\theta_{dk}) = \mathbf{x}_k^T \boldsymbol{\beta}_d$, $k = 1, ..., K$; $d = 1, ..., D$
  - (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 $\{F_1, ..., F_K\}$, though a weaker type of dependence than dependent DP priors (MacEachern, 2000). [Dependent nonparametric prior models will be studied later in the course.]

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 and Parmigiani, 2002)

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

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

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

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 $x_1 < x_2 < \ldots < x_k$).
  - 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 the dose level, i.e., F is a non-decreasing function, i.e., F can be modeled as a c.d.f. on $\mathcal{X} \subseteq \mathbb{R}$.

Dose-response modeling with Dirichlet process priors

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

- Parametric modeling: F is assumed to be a member of a parametric family of c.d.f.s (e.g., logit or probit models).

- Bayesian nonparametric modeling: nonparametric priors for F, i.e., priors for the space of c.d.f.s on $\mathcal{X}$.
  - Work based on a DP prior for F: Antoniak (1974), Bhattacharya (1981), Disch (1981), Kuo (1983), Gelfand and Kuo (1991), Mukhopadhyay (2000).

Dose-response modeling with Dirichlet process priors

- Assuming (conditionally) 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 $p_i = F(x_i)$ for $i = 1, ..., k$.

- If the prior for F is a DP with precision parameter $\alpha > 0$ and centering c.d.f. F_0 (the prior guess for the potency curve), then a priori

$$(p_1, p_2 - p_1, ..., p_k - p_{k-1}, 1 - p_k)$$

follows a Dirichlet distribution with parameters

$$(\alpha F_0(x_1), \alpha(F_0(x_2)-F_0(x_1)), ..., \alpha(F_0(x_k)-F_0(x_{k-1})), \alpha(1-F_0(x_k))).$$

Dose-response modeling with Dirichlet process priors

- The posterior for F is an MDP (Antoniak, 1974).
  - Posterior distribution is difficult to work with analytically; Antoniak (1974) obtained the point estimate when $k = 2$.
  - MCMC techniques enable full inference for the dose-response curve (Gelfand and Kuo, 1991) and for the dose that corresponds to a specified probability of response (Mukhopadhyay, 2000).

- Bioassay modeling with a DP prior for the dose-response curve is an example of semiparametric *isotonic* regression, that is, regression modeling with monotonic regression functions. Further work with DP priors for:
  - continuous response distributions (Lavine and Mockus, 1995)
  - count responses (Farah, Kottas and Morris, 2013).

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

Bayesian nonparametric modeling for cytogenetic dosimetry

- Bayesian nonparametric modeling for polytomous response (Kottas, Branco and Gelfand, 2002).
- Consider simple case with $r = 3 \Rightarrow$ 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 c.d.f.s (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 $\mathbf{q}_\ell = (q_{\ell,1}, ..., q_{\ell,k})$, $\ell = 1, 2$, where $q_{\ell,i} = G_\ell(x_i)$.

Bayesian nonparametric modeling for cytogenetic dosimetry

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

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

  where

$$\gamma_i = \alpha_1 (G_{01}(x_i) - G_{01}(x_{i-1})), \quad \delta_i = \alpha_2 (G_{02}(x_i) - G_{02}(x_{i-1})).$$

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

Bayesian nonparametric modeling for cytogenetic dosimetry

- Inference based on an MCMC algorithm.
  - For any unobserved (but specified) dose level x_0, the posterior distribution for $q_{\ell,0} = G_\ell(x_0)$ for $\ell = 1, 2$, is given by

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

    where $p(q_{\ell,0} \mid \mathbf{q}_\ell)$ is a rescaled Beta distribution.
  - The inversion problem (inference for an unknown x_0 for specified response values $\mathbf{y}_0 = (y_{01}, y_{02}, y_{03})$) can be handled by extending the MCMC algorithm 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,

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

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

Simulated data examples

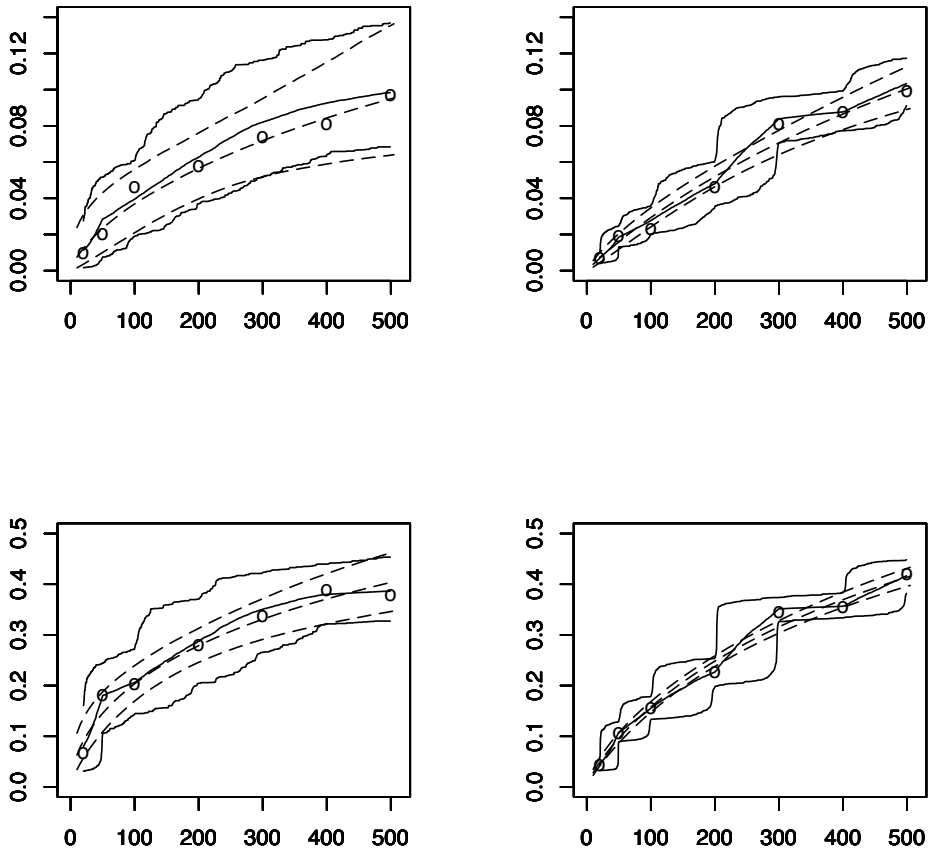


Figure 1.4: Two data sets generated from the parametric model. Posterior inference for F_1 (upper panels) and F_2 (lower panels) under the parametric (dashed lines) and nonparametric (solid lines) model. "o" denotes the observed data. The left and right panels correspond to the data set with the smaller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

Simulated data examples

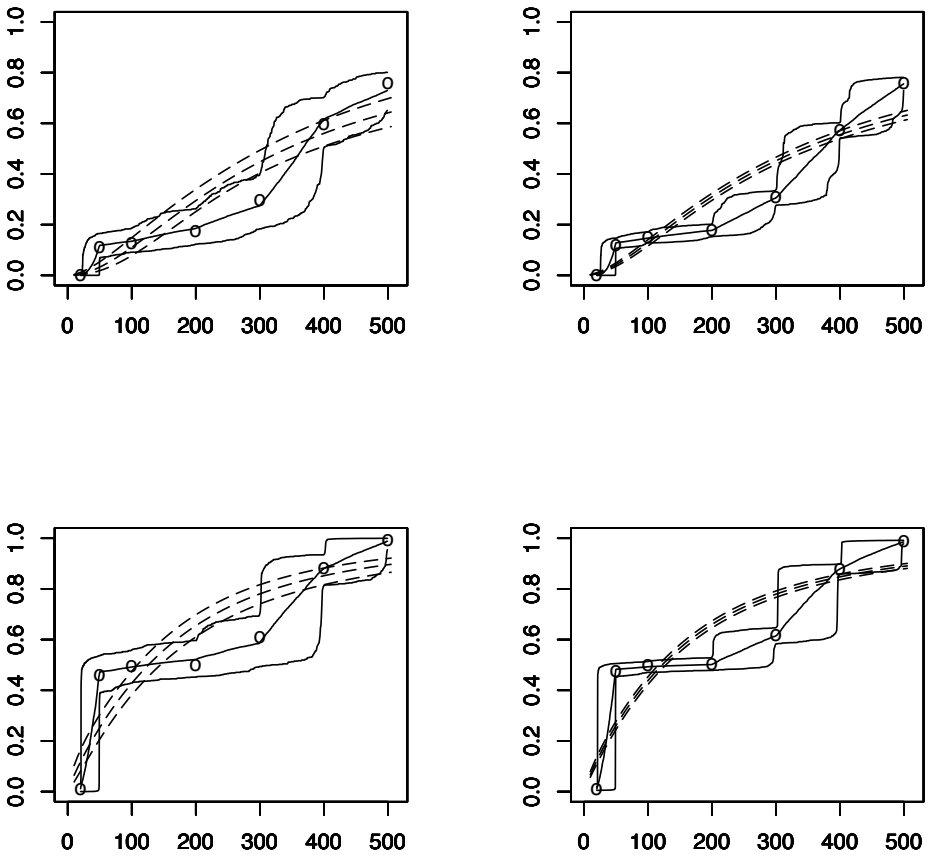


Figure 1.5: Two data sets generated using non-standard (bimodal) shapes for F_1 and F_2. Posterior inference for F_1 (upper panels) and F_2 (lower panels) under the parametric (dashed lines) and nonparametric (solid lines) model. "o" denotes the observed data. The left and right panels correspond to the data set with the smaller and large sample sizes, respectively.