ISBA 2021 short course: Applied Bayesian Nonparametric Mixture Modeling

Athanasios Kottas

Department of Statistics, University of California, Santa Cruz

ISBA 2021 World Meeting
Virtual Conference
Saturday June 26, 2021

Logistics

- Course notes (and references) are available from:

  `https://users.soe.ucsc.edu/~thanos`

- Many thanks to UCSC grad students:
  - Jizhou Kang
  - Hyotae Kim
  - Chunyi Zhao
  - Xiaotian Zheng

  who will help with monitoring questions in the zoom chat.

  (Xiaotian is presenting in Session C09; Jizhou, Hyotae and Chunyi are presenting in Session C20)

Course objectives

- To provide an introduction to Bayesian nonparametric methods, with emphasis on modeling approaches built from nonparametric mixtures.
  - Focus on ideas, methods, and modeling.
  - Examples drawn from density estimation, nonparametric regression, dose-response modeling, and inference for point processes.
- No previous experience with Bayesian nonparametrics is assumed.
  - We assume background on parametric Bayesian hierarchical modeling and computing.

Outline

1. Bayesian nonparametrics: introduction and motivation; overview of nonparametric priors for spaces of random functions.

2. The Dirichlet process as a prior for random distributions: definitions; properties; inference.

3. Dirichlet process mixture models: properties; posterior simulation; applications.

4. Nonparametric priors for dependent distributions: Dependent Dirichlet processes; hierarchical nonparametric prior models for finite collections of distributions; spatial Dirichlet processes; applications.

1. Bayesian Nonparametrics: Introduction and Motivation

Parametric vs. nonparametric Bayes: a simple example

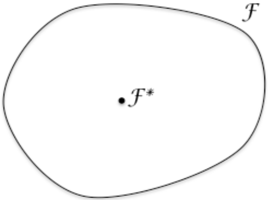
- Let $y_i \mid G \overset{i.i.d.}{\sim} G$, with $G \in \mathcal{F}^*$,

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

- 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

$$\mathcal{F} = \{\text{all distributions on } \mathbb{R}\}.$$

- **Nonparametric Bayes** involves priors on much larger subsets of $\mathcal{F}$, in fact, generally on the entire space $\mathcal{F}$.
- Parametric vs. nonparametric Bayes: finite-dimensional parameter space (e.g., two parameters for $\mathrm{N}(\mu, \tau^2)$) vs. infinite-dimensional space, $\{G(y) : y \in \mathbb{R}\}$.



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 functions, intensity functions, link functions, spectral densities, covariance functions, ...

- Wandering nonparametrically near a standard class.

- More generally, enriching usual parametric models, typically leading to semiparametric models.

- *Bayesian nonparametrics*, an oxymoron? very different from classical nonparametric estimation techniques.

Bayesian vs. classical nonparametrics

- Consider again estimation for a distribution on $\mathbb{R}$.
- Standard classical nonparametric estimates: histogram or empirical distribution function
  - purely data-based estimates
  - no probability model for the underlying data-generating distribution (*nonparametric*)
  - limitations in terms of uncertainty quantification for point estimates, prediction, etc.

- In contrast, Bayesian nonparametric methods build from probability models for the unknown (random) distribution
  - the model for the distribution is not restricted to a parametric family of distributions (*nonparametric*)
  - priors that support the entire space of distributions on $\mathbb{R}$
  - advantages w.r.t. robust inference, uncertainty quantification, prediction, etc., but with more demanding implementation.

Bayesian vs. classical nonparametrics

- Another example for a semiparametric model setting: linear regression with unknown error distribution
  - continuous (real-valued) responses y_i with covariate vector $\boldsymbol{x}_i$

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \overset{i.i.d.}{\sim} G$$

  where G is the error distribution.

- Least-squares is a classical semiparametric estimation technique: it provides estimates for the regression coefficients $\boldsymbol{\beta}$ *without* assuming a probability model for the error distribution.

- In contrast, a Bayesian semiparametric modeling approach would proceed with a parametric prior for $\boldsymbol{\beta}$ and a nonparametric prior for G, where now the space of interest involves all distributions on $\mathbb{R}$ with zero mean (or median or mode).

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 ...) $\rightarrow$ 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 $\{G(\cdot \mid \theta) : \theta \in \Theta\}$ → 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

- Review papers on Bayesian nonparametrics:
  - Sinha & Dey (1997); Gelfand (1999); Walker, Damien, Laud & Smith (1999); Müller & Quintana (2004); Hanson, Branscum & Johnson (2005); Müller & Mitra (2013); Müller, Quintana & Page (2018).
- Books/edited volumes:
  - Dey, Müller & Sinha (1998); Ghosh & Ramamoorthi (2003); Hjort, Holmes, Müller & Walker (2010); Müller & Rodriguez (2013); Phadia (2013); Mitra & Müller (2015); Müller, Quintana, Jara & Hanson (2015); Ghosal & van der Vaart (2017).
- Software:
  - The `DPpackage` in `R` (archived): `https://cran.r-project.org/web/packages/DPpackage/index.html`
  - `NIMBLE`: `https://r-nimble.org/`

Methods for construction of NPB models for distributions

- The object is to define priors over spaces of distributions on a sample space $\mathcal{X}$; say, $\mathcal{X} = \mathbb{R}$ (although the space can be more general).

- Methods for constructing nonparametric priors for distributions:
  - Random probability measures
  - Neutral to the right processes
  - Tailfree processes (Pólya tree priors)
  - Constructions through exchangeable sequences
  - Normalized random measures with independent increments
  - Countable representations for random discrete distributions
  - Nonparametric mixture models

Constructing NPB priors through stochastic processes

- A natural way to conceptualize a nonparametric prior is through a stochastic process with sample paths that are distribution functions or, more generally, distributions (probability measures) on the sample space of interest, $\mathcal{X}$.

- Let $\mathcal{X} = \mathbb{R}$, and focus on space $\mathcal{G}_F = \{$distribution functions F on $\mathbb{R}\}$.

- Then, we can generate random distribution functions on $\mathbb{R}$ through a stochastic process $\{F(\omega, t) : \omega \in \Omega,\ t \in \mathbb{R}\}$, with $\mathbb{R}$ for the index set, and $[0, 1]$ for the state space, such that
  - for any fixed $\omega \in \Omega$, $F_\omega(\cdot) \equiv F(\omega, \cdot) : \mathbb{R} \rightarrow [0, 1]$ is a distribution function on $\mathbb{R}$ (the sample paths are distribution functions)
  - for any fixed $t \in \mathbb{R}$, $F_t(\cdot) \equiv F(\cdot, t) : \Omega \rightarrow [0, 1]$ is a random variable, and for any fixed $t_1, ..., t_k \in \mathbb{R}$, $(F_{t_1}, ..., F_{t_k})$ is a random vector.

Constructing NPB priors through stochastic processes

- A key example of the stochastic process approach for distribution functions is the class of neutral to the right (NTTR) process priors (Doksum, 1974; Ferguson and Phadia, 1979; Damien et al., 1995)
  - Usually applied to distribution functions on $\mathbb{R}^+$ (applications in survival and reliability analysis).
  - Write the random distribution function as $F(t) = 1 - \exp(-Z(t))$, where $\mathcal{Z} = \{Z(t) : t \in \mathbb{R}^+\}$ is a **NTTR Lévy process** (here, we are skipping ω from the notation for the sample paths).
  - By definition, a NTTR Lévy process $\mathcal{Z}$ has non-negative, independent increments, and is, almost surely, non-decreasing, right continuous, with $\lim_{t \to 0} Z(t) = 0$ and $\lim_{t \to \infty} Z(t) = \infty$ (the sample paths of $\mathcal{Z}$ have at most countably many fixed points of discontinuity).
  - Since $Z(t) = -\log(1 - F(t))$, in the context of survival analysis, the prior for the distribution function is induced by a prior for the cumulative hazard function.

Constructing NPB priors through stochastic processes

- Working with distribution functions for the stochastic process sample paths is practical for distributions defined on (subsets of) $\mathbb{R}$.

- In full generality, we would like to define the prior as a stochastic process with sample paths that are distributions on $(\mathcal{X}, \mathcal{B})$.
    - Technical: $\mathcal{B}$ denotes the σ-field of measurable subsets of $\mathcal{X}$ (e.g., the Borel σ-field for $\mathcal{X} \subseteq \mathbb{R}^d$).

- Again, let $\mathcal{X} = \mathbb{R}$ (although the sample space can now be much more general), and now focus on space $\mathcal{G}_Q = \{\text{distributions } Q \text{ on } \mathbb{R}\}$.

Constructing NPB priors through stochastic processes

- Now, we seek to generate random distributions (random probability measures) on $\mathbb{R}$ through a stochastic process $\{Q(\omega, B) : \omega \in \Omega,\ B \in \mathcal{B}\}$, with $\mathcal{B}$ for the index set, and $[0,1]$ for the state space, such that
  - for any fixed $\omega \in \Omega$, $Q_\omega(\cdot) \equiv Q(\omega, \cdot) : \mathcal{B} \to [0,1]$ is a probability measure (distribution) on $\mathbb{R}$ (the sample paths are distributions)
  - for any fixed $B \in \mathcal{B}$, $Q(B) : \Omega \to [0,1]$ is a random variable, and for any fixed $B_1, ..., B_k \in \mathcal{B}$, $(Q(B_1), ..., Q(B_k))$ is a random vector (switching to notation $Q(B)$ for random variable Q_B).

- To make the stochastic process approach for distributions operational, we need finite dimensional distributions (f.d.d.s) for the random vector $(Q(B_1), ..., Q(B_k))$, which satisfy appropriate consistency conditions for existence of the stochastic process.

- f.d.d.s for collections of random probabilities? how about Dirichlet f.d.d.s? → **Dirichlet process**!

2. The Dirichlet Process

The Dirichlet process as a model for random distributions

- 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 original DP definition (Ferguson, 1973) involves a stochastic process (random probability measure) that generates distributions (probability measures) on $\mathcal{X}$, and thus, for $\mathcal{X} \subseteq \mathbb{R}^d$, it also generates distribution functions on $\mathcal{X}$.

Consistency conditions

- Defining the prior as a stochastic process with sample paths that are distributions on a particular sample space $\mathcal{X}$.

- 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 $(Q(B_{1,1}), \ldots, Q(B_{m,k}))$ for each finite collection $B_{1,1}, \ldots, B_{m,k}$ 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
  - $(Q(\cup_{i=1}^{k} B_{1,i}), \ldots, Q(\cup_{i=1}^{k} B_{m,i}))$ and $(\sum_{i=1}^{k} Q(B_{1,i}), \ldots, \sum_{i=1}^{k} Q(B_{m,i}))$ are equal in distribution

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

Motivating the construction of the Dirichlet process

- Consider a sample space with only two outcomes, $\mathcal{X} = \{0, 1\}$, such that defining a distribution on $\mathcal{X}$ requires only one probability, x.
  - A natural prior for x is the Beta distribution.

- More generally, if $\mathcal{X}$ is finite with q elements, the distribution is given by a probability vector, $(x_1, \ldots, x_q)$, i.e., $x_i \geq 0$ with $\sum_{i=1}^{q} x_i = 1$.
  - Now, the natural prior for $(x_1, \ldots, x_q)$ is the Dirichlet distribution.

- To handle uncountable spaces, such as $\mathcal{X} = \mathbb{R}$, consider finite collections of (measurable) subsets of $\mathcal{X}$, say, $B_1, ..., B_k$, with the extra structure that they form a partition of $\mathcal{X}$.
  - The Dirichlet distribution is a natural candidate for the distribution of the probability vector $(Q(B_1), ..., Q(B_k))$.
  - But care is needed, a system of Dirichlet f.d.d.s must be consistent with *any* other partition (any finite k and any collection $(B_1, ..., B_k)$).
  - The Dirichlet distribution works with an appropriate choice for its parameter vector (the key reason is an additivity property which arises from the additivity of the gamma distribution).

Definition/properties of the Dirichlet distribution

- Start with independent random variables

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

with $a_j > 0$.

- Define

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

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

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

Definition/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:

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

$$\mathrm{E}(Y_iY_j)=\frac{a_ia_j}{\sum_{\ell=1}^{k} a_\ell(1+\sum_{\ell=1}^{k} a_\ell)}, \quad \text{for} \;\; i\neq j.$$

- 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:
  - α → a positive scalar parameter;
  - Q_0 → a specified probability measure (distribution) on $(\mathcal{X}, \mathcal{B})$.

- **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:
    - if $Z_r \overset{ind.}{\sim} \text{gamma}(a_r, 1)$, $r = 1, ..., N$, then $\sum_{r=1}^{N} Z_r \sim \text{gamma}(\sum_{r=1}^{N} a_r, 1)$.

- 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 $1 \leq m_1 < ... < m_M = k$, 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 $\text{Dirichlet}(\sum_{i=1}^{m_1} a_i, \sum_{i=m_1+1}^{m_2} a_i, ..., \sum_{i=m_{M-1}+1}^{m_M} a_i)$ 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 $Q(B) \sim \mathrm{Beta}(\alpha Q_0(B), \alpha Q_0(B^c))$, and thus

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

- 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.

Support of the Dirichlet process

- The study of support requires a σ-field on the space of all distributions on $\mathcal{X}$ $\rightarrow$ one option: define Borel sets through the topology of weak convergence
  - Take $\mathcal{X} = \mathbb{R}$, such that $\mathcal{G}_Q = \{\text{distributions } Q \text{ on } \mathbb{R}\}$.
  - Def.: $Q_n \rightarrow Q$ weakly as $n \rightarrow \infty$ if $Q_n((-\infty, x]) \rightarrow Q((-\infty, x])$ as $n \rightarrow \infty$, for all $x \in \mathbb{R}$ such that $Q(\{x\}) = 0$ (i.e., weak convergence for distributions on $\mathbb{R}$ corresponds to convergence in distribution).
- Ferguson (1973) 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.
  - If μ and λ are two probability measures on the same space Ω, λ is absolutely continuous w.r.t. μ (notation, $\lambda << \mu$) if-f $\lambda(A) = 0$ for any measurable $A \subseteq \Omega$ such that $\mu(A) = 0$.
  - Proposition 3 (Ferguson, 1973). Consider a DP on $(\mathcal{X}, \mathcal{B})$ with base distribution Q_0, and S a fixed distribution on $(\mathcal{X}, \mathcal{B})$ with $S << Q_0$. Then, for any positive integer m, any measurable sets $B_1, ..., B_m$, and any $\varepsilon > 0$, $\mathcal{P}(|Q(B_i) - S(B_i)| < \varepsilon, \text{ for } i = 1, ..., m) > 0$ (where $\mathcal{P}$ denotes the DP random probability measure).
- Practical implication: Q_0 must have (at least) the same support with the distributions modeled with the DP prior.

Random c.d.f.s from 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, where G_0 is a specific 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

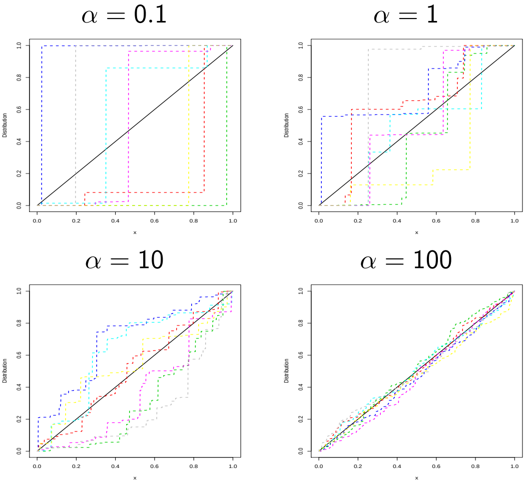
$$(G(x_1), G(x_2) - G(x_1), ..., G(x_k) - G(x_{k-1}), 1 - G(x_k))$$

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 2.1): $\mathcal{X} = (0,1)$, $G_0(x) = x$, $x \in (0,1)$ (Unif$(0,1)$ centering distribution).

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



$\alpha = 0.1$

$\alpha = 1$

$\alpha = 10$

$\alpha = 100$

Figure 2.1: 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).

- Chapter 4 of Ghosal and van der Vaart (2017) provides a detailed account of several properties of the Dirichlet process.

Constructive definition of the DP

- Due to Sethuraman and Tiwari (1982) and Sethuraman (1994).

- Let $\{z_r : r = 1, 2, \ldots\}$ and $\{\vartheta_\ell : \ell = 1, 2, \ldots\}$ be independent sequences of i.i.d. random variables
  - $z_r \overset{i.i.d.}{\sim}$ Beta$(1, \alpha)$, $r = 1, 2, \ldots$.
  - $\vartheta_\ell \overset{i.i.d.}{\sim} G_0$, $\ell = 1, 2, \ldots$.

- Define $\omega_1 = z_1$ and $\omega_\ell = z_\ell \prod_{r=1}^{\ell-1}(1 - z_r)$, for $\ell = 2, 3, \ldots$.

- Then, a realization G from DP(α, G_0) is (almost surely) of the form

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

where δ_a 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 a *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

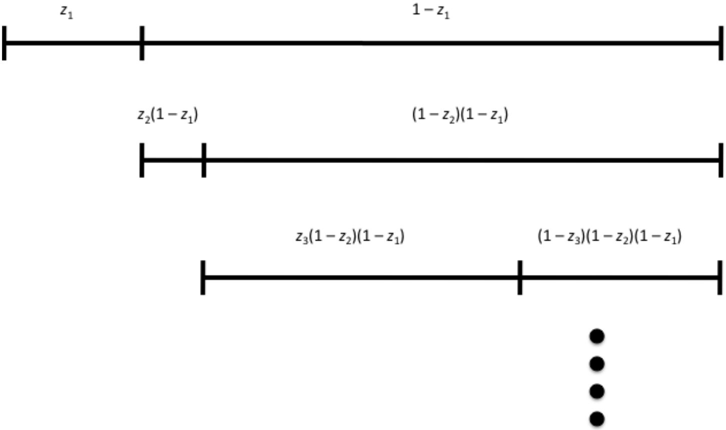
- Start with a stick of length 1 (representing the total probability to be distributed among the different atoms).

- 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 ....

The stick-breaking construction



The stick-breaking construction

- The random series $\sum_{\ell=1}^{\infty} \omega_\ell$ converges almost surely to 1:
  - note that $\sum_{\ell=1}^{n} \omega_\ell = 1 - U_n$, where $U_n = \prod_{r=1}^{n}(1 - z_r)$
  - using the independence of the z_r, and their $\text{Beta}(1, \alpha)$ distribution,

$$\mathrm{E}(U_n) = \mathrm{E}\{\prod_{r=1}^{n}(1 - z_r)\} = \prod_{r=1}^{n} \mathrm{E}(1 - z_r) = \{\alpha/(\alpha + 1)\}^n$$

  - therefore, $\lim_{n \to \infty} \mathrm{E}(U_n) = 0$, i.e., $\{U_n : n \geq 1\}$ converges to 0 in mean of order 1, which implies that $\{U_n : n \geq 1\}$ converges to 0 in probability
  - moreover, $\{U_n : n \geq 1\}$ is an (almost surely) decreasing sequence of positive random variables, and thus its convergence in probability to 0 implies almost sure convergence to 0
  - therefore, the sequence of partial sums, $\{\sum_{\ell=1}^{n} \omega_\ell : n \geq 1\}$, converges almost surely to 1 as $n \to \infty$, which, by definition, implies that the random series $\sum_{\ell=1}^{\infty} \omega_\ell$ converges almost surely to 1.

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 $j = 1, \dots, J-1$, and $p_J = 1 - \sum_{j=1}^{J-1} \omega_j = \prod_{r=1}^{J-1}(1 - z_r)$.

- To specify J, 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

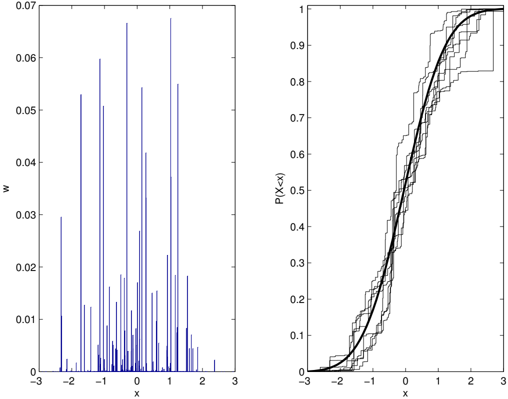


Figure 2.2: Illustration for a DP with $G_0 = \mathrm{N}(0, 1)$ and $\alpha = 20$. In the left panel, the spiked lines are located at 1000 N(0, 1) draws with heights given by the (truncated) stick-breaking weights. These spikes are then summed to generate one c.d.f. sample path. The right panel shows 8 such sample paths indicated by the lighter jagged lines. The heavy smooth line indicates the N(0, 1) c.d.f.

DP properties using the constructive definition

- **Mean functional.** Consider $\mathcal{X} = \mathbb{R}$, and assume $G \sim \mathrm{DP}(\alpha, G_0)$ such that G_0 has finite mean, i.e., $\int |x| \, \mathrm{d}G_0(x) < \infty$. Then, $\mu(G) = \int x \, \mathrm{d}G(x)$ is an (almost surely) finite r.v., and $\mathrm{E}(\mu(G)) = \mu(G_0) = \int x \, \mathrm{d}G_0(x)$.

  - Let $\mu^*(G) = \int |x| \, \mathrm{d}G(x)$. Then, using the DP constructive definition, $G = \sum_{\ell=1}^{\infty} \omega_\ell \, \delta_{\vartheta_\ell}$, and the monotone convergence theorem (MCT),

    $$\mathrm{E}(\mu^*(G)) = \mathrm{E}(\sum_{\ell=1}^{\infty} \omega_\ell \, |\vartheta_\ell|) = \sum_{\ell=1}^{\infty} \mathrm{E}(\omega_\ell \, |\vartheta_\ell|) = \sum_{\ell=1}^{\infty} \mathrm{E}(\omega_\ell) \, \mathrm{E}(|\vartheta_\ell|).$$

  - Now, $\mathrm{E}(|\vartheta_\ell|)$ is a finite constant that does not depend on ℓ (say, C), since the ϑ_ℓ are i.i.d. G_0, and G_0 is assumed to have finite mean. Therefore, $\mathrm{E}(\mu^*(G)) = C \sum_{\ell=1}^{\infty} \mathrm{E}(\omega_\ell) = C < \infty$, using again MCT.
  - Since $\mu^*(G)$ is a positive valued r.v. with finite mean, we have that $\mu^*(G)$ is finite almost surely, and thus $\mu(G)$ is absolutely convergent almost surely.

  - Now, $\mu(G) = \int x \, \mathrm{d}G(x) = \sum_{\ell=1}^{\infty} \omega_\ell \, \vartheta_\ell \leq \mu^*(G)$, with $\mathrm{E}(\mu^*(G)) < \infty$, and thus from the dominated convergence theorem (DCT)

    $$\mathrm{E}(\mu(G)) = \sum_{\ell=1}^{\infty} \mathrm{E}(\omega_\ell \, \vartheta_\ell) = \sum_{\ell=1}^{\infty} \mathrm{E}(\omega_\ell) \, \mathrm{E}(\vartheta_\ell) = \mu(G_0)$$

    using MCT for the last equation.

DP properties using the constructive definition

- Analogously, we can obtain the prior expectation for the variance functional, $\sigma^2(G) = \int x^2 \, \mathrm{d}G(x) - (\int x \, \mathrm{d}G(x))^2$, and the prior variance for the mean functional. For such results, we need G_0 to have finite first and second moments.

- Consider $\mathcal{X} = \mathbb{R}$, and assume $G \sim \mathrm{DP}(\alpha, G_0)$, with $G_0 = \mathrm{N}(0,1)$.
  - Using MCT and DCT, and the expression for $\mathrm{E}(\omega_\ell^2)$ (readily available from the indepence of the beta r.v.s used to define the DP weights), it can be shown that $\mathrm{E}(\int x^2 \, \mathrm{d}G(x)) = 1$, and $\mathrm{E}\{(\int x \, \mathrm{d}G(x))^2\} = 1/(\alpha+1)$.
  - Therefore, $\mathrm{E}(\sigma^2(G)) = \alpha/(\alpha+1)$.
  - Similarly, $\mathrm{Var}(\mu(G)) = 1/(\alpha+1)$.

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 $z_r \overset{i.i.d.}{\sim} \text{Beta}(a, b)$.
- If $z_r \overset{ind.}{\sim} \text{Beta}(1 - a, b + ra)$, 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 & Yor, 1997).
- The general case, $z_r \overset{i.i.d.}{\sim} \text{Beta}(a_r, b_r)$ (Ishwaran & James, 2001).
- The probit stick-breaking process: $z_r = \Phi(x_r)$, where $x_r \sim N(\mu, \sigma^2)$ and Φ is the standard normal c.d.f. (Rodríguez & 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 & Tardella, 1998), generalized DPs (Hjort, 2000); general stick-breaking priors (Ishwaran & James, 2001).
- Dependent DP priors (MacEachern, 1999, 2000).
- 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

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

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

- That is, the sequence of the X_i follows a generalized Pólya urn scheme such that:
    - $X_1 \sim G_0$, and
    - for any $i = 2, \ldots, n$, $X_i \mid X_1 = x_1, \ldots, X_{i-1} = x_{i-1}$ follows a distribution that places point mass $(\alpha + i - 1)^{-1}$ at x_j, for $j = 1, \ldots, i - 1$, and the remaining 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*.

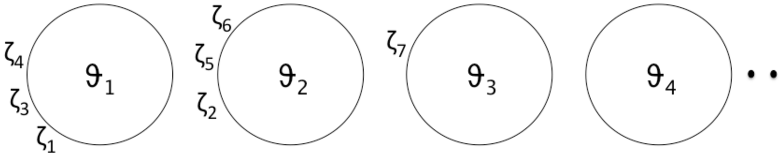
- A sequence of r.v.s, $\{X_n : n \geq 1\}$, (w.l.o.g. on $\mathbb{R}$) is a Pólya sequence with parameters G_0 (a distribution on $\mathbb{R}$) and α (a positive scalar parameter) if for any measurable $B \subset \mathbb{R}$, $\Pr(X_1 \in B) = G_0(B)$, and $\Pr(X_{n+1} \in B \mid X_1, ..., X_n) = (\alpha + n)^{-1}\{\alpha G_0(B) + \sum_{i=1}^{n} \delta_{X_i}(B)\}$ (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 $n \rightarrow \infty$) to a discrete distribution G
  - $G \sim \mathrm{DP}(\alpha, G_0)$
  - $X_1, X_2, ... \mid G$ 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.



Exchangeability and Nonparametric Bayes

- *de Finetti's representation theorem* provides an interesting connection between exchangeability and nonparametric priors on spaces of distributions. Below is an overview focusing on distributions on $\mathcal{X} = \mathbb{R}$.
- Def.: Random variables $X_1, ..., X_n$ are (finitely) exchangeable if their joint distribution is invariant to permutations of the r.v. indexes, i.e., $p(x_1, ..., x_n) = p(x_{\pi(1)}, ..., x_{\pi(n)})$, for any permutation π of $\{1, ..., n\}$. A countable collection of r.v.s is (infinitely) exchangeable if the condition above holds true for every finite subset of its r.v.s.
- **Representation theorem for binary r.v.s.** Consider an exchangeable sequence of binary 0/1 r.v.s $\{X_i : i = 1, 2, ...\}$. Then, there exists a distribution (c.d.f.) G on $(0, 1)$ such that for any n and any $(x_1, ..., x_n)$:

$$p(x_1, ..., x_n) = \int_0^1 \left\{ \prod_{i=1}^{n} \theta^{x_i} (1 - \theta)^{1 - x_i} \right\} \mathrm{d}G(\theta)$$

  - Hence, for any n, the joint distribution of $X_1, ..., X_n$ can be obtained by generating a probability θ from distribution G, and then taking $X_1, ..., X_n \mid \theta \overset{i.i.d.}{\sim} \text{Bernoulli}(\theta)$.

Exchangeability and Nonparametric Bayes

- **de Finetti's representation theorem.** Consider an exchangeable sequence of $\mathbb{R}$-valued r.v.s $\{X_i : i = 1, 2, ...\}$ with joint distribution P. Then, there exists a random probability measure $\mathcal{P}$ on the space of distributions on $\mathbb{R}$ such that for any n and any (measurable) sets $(B_1, ..., B_n)$:

$$P(X_1 \in B_1, ..., X_n \in B_n) = \int \left\{ \prod_{i=1}^{n} G(B_i) \right\} \mathrm{d}\mathcal{P}(G)$$

  - Hence, for any n, the joint distribution of $X_1, ..., X_n$ can be obtained by selecting $G \sim \mathcal{P}$, and then taking $X_1, ..., X_n \mid G \overset{i.i.d.}{\sim} G$.
  - $\mathcal{P}$ is the de Finetti measure for the exchangeable sequence. Given the joint distribution of the X_i, the de Finetti measure is unique.

- The generalized Pólya sequence

$$X_1 \sim G_0, \quad X_{n+1} \mid X_1, ..., X_n \sim \frac{\alpha G_0 + \sum_{i=1}^{n} \delta_{X_i}}{\alpha + n}$$

can be verified to be exchangeable. Therefore, the $\mathrm{DP}(\alpha, G_0)$ is the de Finetti measure for this exchangeable sequence.

Prior to posterior updating with DP priors

- The DP is a conjugate prior under i.i.d. sampling.
- Assume data $y_i \mid G \overset{i.i.d.}{\sim} G$, for $i = 1, ..., n$, and $G \sim \mathrm{DP}(\alpha, G_0)$. Then, the posterior distribution of G is the $\mathrm{DP}(\tilde{\alpha}, \tilde{G}_0)$, where

$$\tilde{\alpha} = \alpha + n, \qquad \tilde{G}_0 = \frac{\alpha G_0 + \sum_{i=1}^{n} \delta_{y_i}}{\alpha + n}$$

  (Theorem 1, Ferguson (1973))

- For $\mathcal{X} = \mathbb{R}$, the c.d.f. associated with $\tilde{G}_0$ is

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

- 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

- Sketch of the proof for $\mathcal{X} = \mathbb{R}$.
  - Discretize the space through fixed grid $x_1 < \ldots < x_K$ (K can be arbitrarily large).
  - The random c.d.f. G is represented (approximated) by parameter vector $(G(x_1), \ldots, G(x_K))$. Equivalently, the random distribution G is represented in terms of its probabilities on sets $B_0 = (-\infty, x_1]$, $B_{k-1} = (x_{k-1}, x_k]$, for $k = 2, \ldots, K$, $B_K = (x_K, \infty)$.
  - From the DP definition, $(G(x_1), G(x_2) - G(x_1), \ldots, G(x_K) - G(x_{K-1}), 1 - G(x_K)) \sim$ Dirichlet$(\alpha G_0(x_1), \alpha(G_0(x_2) - G_0(x_1)), \ldots, \alpha(G_0(x_K) - G_0(x_{K-1})), \alpha(1 - G_0(x_K)))$.
  - Consider $n = 1$. Then, the "likelihood" for the single observation y is multinomial, so up to proportionality constant:

    $$(G(x_1))^{\delta_y(B_0)} (G(x_2) - G(x_1))^{\delta_y(B_1)} \ldots (G(x_K) - G(x_{K-1}))^{\delta_y(B_{K-1})} (1 - G(x_K))^{\delta_y(B_K)}$$

    where $\delta_y(B_k) = 1$ if $y \in B_k$ (and $\delta_y(B_k) = 0$, otherwise), for $k = 0, 1, \ldots, K$.
  - Therefore, the posterior distribution for $(G(x_1), G(x_2) - G(x_1), \ldots, G(x_K) - G(x_{K-1}), 1 - G(x_K))$ is Dirichlet with updated parameters $(\alpha G_0(x_1) + \delta_y(B_0), \alpha(G_0(x_2) - G_0(x_1)) + \delta_y(B_1), \ldots, \alpha(G_0(x_K) - G_0(x_{K-1})) + \delta_y(B_{K-1}), \alpha(1 - G_0(x_K)) + \delta_y(B_K))$.
  - Using the DP definition through its Dirichlet f.d.d.s, we conclude that $G \mid y$ follows a DP with updated finite measure $\alpha G_0 + \delta_y$, equivalently, with precision parameter $\alpha + 1$ and centering distribution $\alpha(\alpha + 1)^{-1} G_0 + (\alpha + 1)^{-1} \delta_y$.
  - The result can be extended to any n by induction.

Prior to posterior updating with DP priors

- For $\mathcal{X} = \mathbb{R}$, the posterior mean estimate for the random c.d.f. at any point y, $G(y)$, is given by:

$$\mathrm{E}\left\{G(y) \mid y_1, \ldots, y_n\right\} = \frac{\alpha}{\alpha + n} G_0(y) + \frac{n}{\alpha + n} G_n(y)$$

where $G_n(y) = n^{-1} \sum_{i=1}^{n} 1_{[y_i,\infty)}(y)$ 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 does **not** correspond to a "noninformative" DP prior specification; recall that α controls both the variance and the extent of discreteness for the DP prior.

Prior to posterior updating with DP priors

- How about posterior variability?
- For any measurable set B, the posterior mean for $G(B)$

$$\tilde{G}_0(B) = \mathrm{E}\{G(B) \mid y_1, ..., y_n\} = \frac{\alpha G_0(B) + \sum_{i=1}^{n} \delta_{y_i}(B)}{\alpha + n}$$

and the posterior variance

$$\mathrm{Var}\{G(B) \mid y_1, ..., y_n\} = \frac{\tilde{G}_0(B)(1 - \tilde{G}_0(B))}{1 + \alpha + n} \leq \frac{1}{4(1 + \alpha + n)}$$

  - as $n \to \infty$, the posterior distribution for the random probability $G(B)$ contracts to its mean.

C.d.f. estimation using DP priors

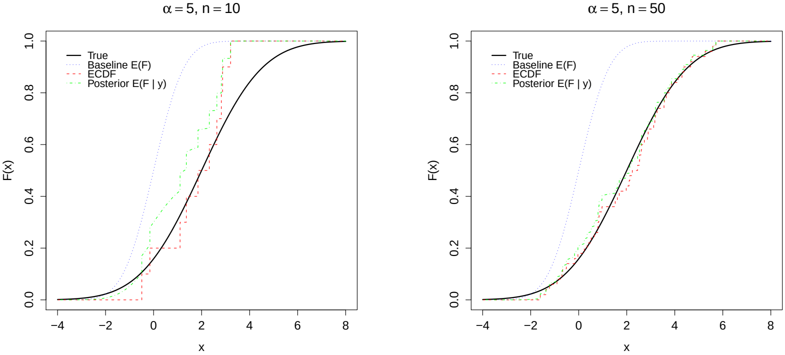


Figure 2.3: Estimating the distribution function under a DP prior, using simulated data. Both the true distribution generating the data and the baseline distribution are Gaussian. The left panel corresponds to a sample of $n = 10$ observations while the right panel corresponds to a sample of $n = 50$ observations.

The DP prediction rule (Pólya urn scheme)

- Having obtained the posterior distribution for G, it's easy to derive the generalized Pólya urn scheme through the posterior predictive distribution.

- Start with $X_i \mid G \overset{i.i.d.}{\sim} G$, for $i = 1, ..., n$, and $G \sim \mathrm{DP}(\alpha, G_0)$.
  - What is the distribution of X_{n+1} given $X_1, ..., X_n$? In the context of Bayesian inference, this is the posterior predictive distribution (so, $X_1, ..., X_n$ represent the r.v.s for the observables in the sample).

- For any measurable set B,

$$p(X_{n+1} \in B, G \mid X_1, ..., X_n) = G(B)\, p(G \mid X_1, ..., X_n)$$

and therefore marginalizing G over its posterior distribution

$$\Pr(X_{n+1} \in B \mid X_1, ..., X_n) = \mathrm{E}(G(B) \mid X_1, ..., X_n) = \frac{\alpha G_0(B) + \sum_{i=1}^{n} \delta_{X_i}(B)}{\alpha + n}$$

  - This is the generalized Pólya conditional distribution, for any $n \geq 1$.
  - For the first member of the sequence, note that $X \mid G \sim G$ and $G \sim \mathrm{DP}(\alpha, G_0)$ implies the marginal $\Pr(X \in B) = \int \Pr(X \in B \mid G)\, \mathrm{d}\mathcal{P}(G) = \int G(B)\, \mathrm{d}\mathcal{P}(G) = \mathrm{E}(G(B)) = G_0(B)$.

Connection with the Bayesian Bootstrap

- Consider data $= \{y_1, \ldots, y_n\}$ assumed to arise (conditionally) i.i.d. from distribution G (say, on $\mathbb{R}$).
- Bootstrap (Efron, 1979): $G \approx n^{-1} \sum_{i=1}^{n} \delta_{y_i}$, the empirical distribution
    - one bootstrap replication is a simple random sample of size n obtained with replacement from $\{y_1, \ldots, y_n\}$
    - for any distribution functional h → compute $\hat{h}$, the bootstrap draw based on the bootstrap sample → develop the bootstrap distribution for h by sampling (independently) multiple $\hat{h}$ values.
- Bayesian Bootstrap (BB) (Rubin, 1981): same idea but with weighted re-sampling from the data → replace the empirical distribution with $\sum_{i=1}^{n} w_i \delta_{y_i}$, where $(w_1, \ldots, w_n) \sim \text{Dirichlet}(1, \ldots, 1)$.
- DP-based model: $y_i \mid G \overset{i.i.d.}{\sim} G$, for $i = 1, \ldots, n$, with $G \sim \text{DP}(\alpha, G_0)$
    - The weak limit of the posterior distribution $\text{DP}(\tilde{\alpha}, \tilde{G}_0)$, as $\alpha \to 0$, is a DP centered on the finite measure $\sum_{i=1}^{n} \delta_{y_i}$ (so, the precision parameter is n and the centering distribution is $n^{-1} \sum_{i=1}^{n} \delta_{y_i}$) → the distributions it generates are supported on the data points $\{y_1, \ldots, y_n\}$ → this is essentially the BB distribution $\sum_{i=1}^{n} w_i \delta_{y_i}$

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:

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

  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 $G \sim \mathrm{DP}(\alpha, G_0)$.
  - 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:

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

- 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:

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

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

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

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

  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 $\boldsymbol{x}_i$ 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 and Regazzini, 1978) for the cell-specific random distributions F_k, $k = 1, \dots, K$:

    - conditionally on hyperparameters α_k and $\boldsymbol{\theta}_k$, the F_k are assigned independent $\text{DP}(\alpha_k, F_{0k}(\cdot \mid \boldsymbol{\theta}_k))$ priors, where, in general, $\boldsymbol{\theta}_k = (\theta_{1k}, \dots, \theta_{Dk})$

    - 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, $d = 0, 1, \dots, D$)
    - $h_0(\alpha_k) = \boldsymbol{x}_k^T \boldsymbol{\gamma}$, $k = 1, \dots, K$
    - $h_d(\theta_{dk}) = \boldsymbol{x}_k^T \boldsymbol{\beta}_d$, $k = 1, \dots, K$; $d = 1, \dots, D$
    - (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 dependent DP priors (MacEachern, 2000).

Semiparametric regression for categorical responses

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

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

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

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

- Related work for: change-point problems (Mira 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).

3. Dirichlet Process Mixture Models

Motivating Dirichlet process mixtures

- Recall that the Dirichlet process (DP) is a conjugate prior for random distributions under i.i.d. sampling.
- However, posterior draws under a DP model correspond (almost surely) to discrete distributions. This is somewhat unsatisfactory if we are modeling continuous distributions.
- In the spirit of kernel density estimation, one solution is to use convolutions to smooth out posterior estimates.
- In a model-based context, this leads to DP mixture models, i.e., a mixture model where the mixing distribution is unknown and assigned a DP prior (recall that this is different from a mixture of DPs, in which the parameters of the DP are random).
- Strong connection with finite mixture models.
- More generally, we might be interested in using a DP as part of a hierarchical Bayesian model to place a prior on the unknown distribution of some of its parameters (e.g., random effects models). This leads to semiparametric Bayesian models.

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 modeling, and are now relatively easy to implement, using simulation-based model fitting (e.g., Richardson and Green, 1997; Stephens, 2000; Jasra, Holmes and Stephens, 2005).
- Rather than handling the very large number of parameters of finite mixture models with a large number of mixture components, 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.

Finite mixture models

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

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

  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 $i = 1, \dots, n$, given the parameters).

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

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

Finite mixture models

- 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

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

- If we marginalize over L_i, for $i = 1, ..., n$, we recover the original mixture formulation.
- The inclusion of indicator variables is very common in finite mixture models, and it is also used extensively for DP mixtures.

Finite mixture models

- We can also write

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

  where

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

- 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, or more general nonparametric priors for discrete distributions.
- Working with a countable mixture (rather than a finite one) provides theoretical advantages (full support) as well as practical benefits: the number of mixture components is estimated from the data based on a model that supports a countable number of components in the prior.

Definition of the Dirichlet process mixture model

- The **Dirichlet process mixture model**

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

  where $K(y \mid \theta)$ is a parametric distribution function (with parameters θ) on the sample space of interest.

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

- Corresponding mixture density (or probability mass) function,

$$f(y \mid G) = \int k(y \mid \theta) \, \mathrm{d}G(\theta),$$

  where $k(y \mid \theta)$ is the density (or probability mass) function of $K(y \mid \theta)$.

- Because G is random, $F(y \mid G)$ is a random c.d.f., and $f(y \mid G)$ is a random density.

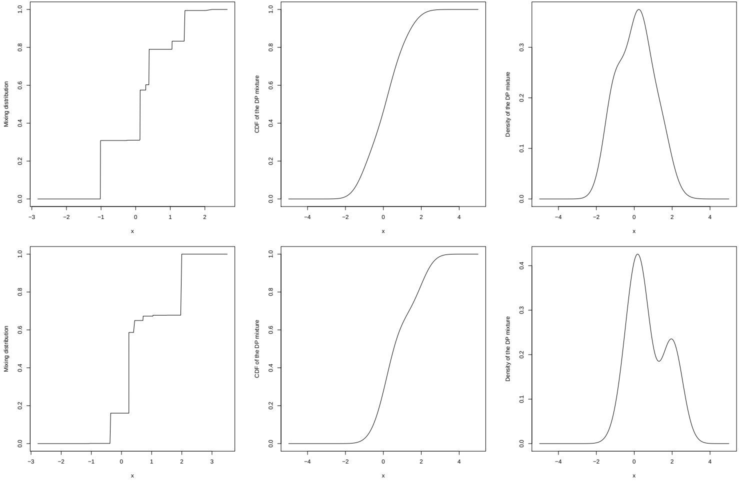


Figure 3.1: Two realizations from a DP($\alpha = 2$, $G_0 = \mathrm{N}(0,1)$) (left column) and the associated cumulative distribution function (center column) and density function (right column) for a location DP mixture of Gaussian kernels with standard deviation 0.6.

An equivalent 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 mixing parameters, and thus makes DP mixture models appealing for many applications, including density estimation and regression.

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

$$f(y \mid G) = \sum_{\ell=1}^{\infty} \omega_\ell \, k(y \mid \vartheta_\ell)$$

  - *Mixture 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$(1, \alpha)$.
  - *Locations*: ϑ_ℓ i.i.d. G_0 (and the sequences $\{z_r : r = 1,2,\ldots\}$ and $\{\vartheta_\ell : \ell = 1,2,\ldots\}$ are independent).

Modeling options

- Contrary to the DP prior, DP mixtures can generate:
  - discrete distributions (e.g., $K(y \mid \theta)$ might be Poisson or binomial)
  - and continuous distributions, either univariate ($K(y \mid \theta)$ can be, e.g., normal, gamma, or uniform) or multivariate (with $K(y \mid \theta)$, say, multivariate normal).

- Much more than density estimation:
  - Non-Gaussian and non-linear regression through DP mixture modeling for the joint response-covariate distribution (*density regression*).
  - Flexible models for ordinal categorical responses.
  - Modeling of point process intensities through density estimation.
  - Time-series and/or spatial modeling, using dependent DP priors for temporally and/or spatially dependent mixing distributions.

Approximation or representation results for mixtures

- (Discrete) normal location-scale mixtures,

$$\sum_{j=1}^{M} w_j \, \mathrm{N}(y \mid \mu_j, \sigma_j^2), \quad y \in \mathbb{R}$$

  can approximate arbitrarily well (as $M \to \infty$) densities on the real line (Ferguson, 1983; Lo, 1984). In fact, the result holds true for mixture kernels from general location-scale families.

- For any non-increasing density $f(t)$ on the positive real line there exists a distribution function G on $\mathbb{R}^+$ such that f can be represented as a scale mixture of uniform densities:

$$f(t) = \int \theta^{-1} 1_{[0,\theta)}(t) \, \mathrm{d}G(\theta), \quad t \in \mathbb{R}^+$$

  - The result yields flexible DP mixture models for symmetric unimodal densities (Brunner and Lo, 1989; Brunner, 1995) as well as general unimodal densities (Brunner, 1992; Lavine and Mockus, 1995; Kottas and Gelfand, 2001; Kottas and Krnjajić, 2009).

Approximation or representation results for mixtures

- Consider a continuous density h on $[0, 1]$, and let H be its c.d.f. Then, the Bernstein density,

$$\sum_{j=1}^{K} \{H(j/K) - H((j-1)/K)\} \, \mathrm{Beta}(u \mid j, K-j+1), \quad u \in [0,1]$$

  converges uniformly to h, as $K \to \infty$.

  - The Bernstein-Dirichlet prior model is based on a DP prior for H (Petrone, 1999a,b).

- Consider a continuous c.d.f. H on $\mathbb{R}^+$. Then, the c.d.f. of the Erlang mixture density

$$\sum_{j=1}^{J} \{H(j\theta) - H((j-1)\theta)\} \, \mathrm{gamma}(t \mid j, \theta), \quad t \in \mathbb{R}^+$$

  converges pointwise to H, as $J \to \infty$ and the scale parameter $\theta \to 0$.

Support of Dirichlet process mixture models

- Results on Kullback-Leibler support for various types of DP mixture models (e.g., Wu and Ghosal, 2008).

- Consider the space of densities defined on sample space $\mathcal{X}$.

- For any density f_0 in that space, the Kullback-Leibler neighborhood of size $\varepsilon > 0$ is given by

$$K_{\varepsilon}(f_0) = \left\{ f : \int f_0(x) \log \left(\frac{f_0(x)}{f(x)} \right) \mathrm{d}x < \varepsilon \right\}$$

- A nonparametric prior model for densities satisfies the Kullback-Leibler property if it assignes positive probability to $K_{\varepsilon}(f_0)$ for any density f_0 in the space of interest, and for any $\varepsilon > 0$ (e.g., Walker, Damien and Lenk, 2004). Typically, several regularity conditions are needed for f_0.

Semiparametric Dirichlet process mixture models

- In many applications, semiparametric DP mixtures are employed

$$y_i \mid G, \phi \overset{i.i.d.}{\sim} f(y_i \mid G, \phi) = \int k(y_i \mid \theta, \phi) \, \mathrm{d}G(\theta), \quad i = 1, \dots, n$$

$$G \sim \mathrm{DP}(\alpha, G_0)$$

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

- For example, semiparametric linear regression model:
  - continuous (real-valued) responses y_i with covariate vector $\boldsymbol{x}_i$

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i; \quad \varepsilon_i \mid G \overset{i.i.d.}{\sim} \int \mathrm{N}(\varepsilon_i \mid 0, \sigma^2) \, \mathrm{d}G(\sigma^2), \;\; G \sim \mathrm{DP}(\alpha, G_0)$$

  - scale normal DP mixture prior for the error distribution; parametric prior for the vector of regression coefficients.

Hierarchical formulation for DP mixture models

- Consider w.l.o.g. the fully nonparametric DP mixture

$$f(y \mid G) = \int k(y \mid \theta)\,\mathrm{d}G(\theta), \qquad G \mid \alpha, \psi \sim \mathrm{DP}(\alpha, G_0(\cdot \mid \psi))$$

- With θ_i a (continuous) latent mixing parameter associated with y_i:

$$\begin{aligned} y_i \mid \theta_i &\overset{ind.}{\sim} k(y_i \mid \theta_i) & i = 1, \dots, n \\ \theta_i \mid G &\overset{i.i.d.}{\sim} G & i = 1, \dots, n \end{aligned}$$

- Alternatively, with discrete latent variables L_i:

$$\begin{aligned} y_i \mid L_i, \{Z_\ell\} &\overset{ind.}{\sim} k(y_i \mid Z_{L_i}) & i = 1, \dots, n \\ L_i \mid \{\omega_\ell\} &\overset{i.i.d.}{\sim} \sum_{\ell=1}^{\infty} \omega_\ell \, \delta_\ell & i = 1, \dots, n \end{aligned}$$

where $\omega_1 = z_1$, $\omega_\ell = z_\ell \prod_{r=1}^{\ell-1}(1 - z_r)$, $\ell \geq 2$, with z_r i.i.d. Beta$(1, \alpha)$, and $Z_\ell \mid \psi \overset{i.i.d.}{\sim} G_0(\cdot \mid \psi)$.

Parametric models in the two limits for α

- Two *limiting* special cases of the DP mixture model.
  - One distinct component, when $\alpha \to 0^+$

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

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

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

- The DP mixture model gives rise to hierarchical structures in between the two parametric *extremes* above.

Connection with finite mixture models

- The countable sum formulation of the DP mixture model has motivated the study of several variants and extensions.
- It also provides 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 and Zarepour, 2000).
- Consider the finite mixture model with J components:

$$\sum_{j=1}^{J} q_j \, k(y \mid \vartheta_j),$$

with $(q_1, \dots, q_J) \sim \text{Dir}(\alpha/J, \dots, \alpha/J)$ and $\vartheta_j \overset{i.i.d.}{\sim} G_0$, $j = 1, \dots, J$.

- When $J \rightarrow \infty$, this model corresponds to a DP mixture with kernel k and a $\text{DP}(\alpha, G_0)$ prior for the mixing distribution.
  - As $J \rightarrow \infty$, $\sum_{j=1}^{J} q_j \delta_{\vartheta_j}$ converges weakly to $\sum_{\ell=1}^{\infty} \omega_\ell \, \delta_{\vartheta_\ell} \sim \text{DP}(\alpha, G_0)$.

Prior specification

- Taking expectation over G with respect to its DP prior $\mathrm{DP}(\alpha, G_0)$, we obtain:

$$\mathrm{E}\{F(y \mid G, \phi)\} = F(y \mid G_0, \phi), \quad \mathrm{E}\{f(y \mid G, \phi)\} = f(y \mid G_0, \phi).$$

- These expressions facilitate prior specification for the parameters ψ of $G_0(\cdot \mid \psi)$.

- On the other hand, recall that for the $\mathrm{DP}(\alpha, G_0)$, α controls how *close* a realization G is to G_0, but also the extent of discreteness of G.

- In the DP mixture model, α controls the prior distribution of the number of distinct elements n^* of vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, and hence the number of distinct mixture components associated with a sample of size n (Antoniak, 1974; Escobar and West, 1995; Liu, 1996).

Pólya urn revisited

- Consider the joint prior distribution for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ that arises from the prior model for the mixing parameters, $\theta_i \mid G \overset{i.i.d.}{\sim} G$ with $G \mid \alpha, \psi \sim \mathrm{DP}(\alpha, G_0(\cdot \mid \psi))$, after integrating G over its DP prior.

- As is essentially always the case for DP mixtures, assume that G_0 is a continuous distribution (i.e., it has no atoms) such that ties can only arise by setting θ_i equal to θ_j, for $j < i$. Denote by g_0 the density function of G_0.

- Using the Pólya urn characterization of the DP,

$$p(\boldsymbol{\theta} \mid \alpha, \psi) = g_0(\theta_1 \mid \psi) \prod_{i=2}^{n} \left\{ \frac{\alpha}{\alpha + i - 1} g_0(\theta_i \mid \psi) + \frac{1}{\alpha + i - 1} \sum_{j=1}^{i-1} \delta_{\theta_j}(\theta_i) \right\}.$$

Pólya urn revisited

- The prior distribution $p(\boldsymbol{\theta} \mid \alpha, \psi)$ can be written in an equivalent form which makes explicit the partitioning (clustering) induced by the discreteness of the DP prior (Antoniak, 1974; Lo, 1984).

- Denote by $\boldsymbol{\pi} = \{s_j : j = 1, ..., n^*\}$ a generic partition of $\{1, ..., n\}$, where: n^* is the number of cells of the partition; n_j is the number of elements in cell s_j; $e_{j,1} < ... < e_{j,n_j}$ are the elements of cell s_j.

- Letting $\mathcal{P}_n$ denote the set of all partitions of $\{1, ..., n\}$,

$$p(\boldsymbol{\theta} \mid \alpha, \psi) = \sum_{\boldsymbol{\pi} \in \mathcal{P}_n} p(\boldsymbol{\pi} \mid \alpha) \left\{ \prod_{j=1}^{n^*} g_0(\theta_{e_{j,1}} \mid \psi) \prod\nolimits_{i=2}^{n_j} \delta_{\theta_{e_{j,1}}}(\theta_{e_{j,i}}) \right\}$$

where $p(\boldsymbol{\pi} \mid \alpha)$ is the DP induced prior probability for partition $\boldsymbol{\pi}$,

$$p(\boldsymbol{\pi} \mid \alpha) = \left(\prod\nolimits_{m=1}^{n} (\alpha + m - 1) \right)^{-1} \alpha^{n^*} \prod_{j=1}^{n^*} (n_j - 1)!$$

Number of distinct components

- Prior expectation and variance for the number of distinct elements (partition cells), $n^* \equiv n^*(n)$, of vector $(\theta_1, \ldots, \theta_n)$.

- Let U_i, for $i = 1, ..., n$, be binary random variables with U_i indicating whether θ_i is a new value drawn from G_0 $(U_i = 1)$ or not $(U_i = 0)$.

- Conditional on α, the U_i are independent Bernoulli random variables with $\Pr(U_i = 1 \mid \alpha) = \alpha/(\alpha + i - 1)$, for $i = 1, ..., n$.

- Since $n^* = \sum_{i=1}^{n} U_i$, we obtain

$$\mathrm{E}(n^* \mid \alpha) = \sum_{i=1}^{n} \frac{\alpha}{\alpha + i - 1} \quad \text{and} \quad \mathrm{Var}(n^* \mid \alpha) = \sum_{i=1}^{n} \frac{\alpha(i-1)}{(\alpha + i - 1)^2}$$

- The prior moments for n^* can be used to guide the choice of the value for α, or the prior parameters for α. (The conditional expectation $\mathrm{E}(n^* \mid \alpha)$ can be averaged over the prior for α to obtain $\mathrm{E}(n^*)$.)

Number of distinct components

- A fairly accurate approximation (for practically all values of n and α):

$$\mathrm{E}(n^* \mid \alpha) \approx \alpha \log\{1 + (n/\alpha)\}.$$

  Hence, $\mathrm{E}(n^* \mid \alpha)$ increases at a logarithmic rate with n (for fixed α).

- Therefore, $\mathrm{E}(n^*(n) \mid \alpha) \to \infty$, as $n \to \infty$. In fact, $n^*(n)$ converges almost surely to ∞, as $n \to \infty$ (Korwar and Hollander, 1973).
  - Even though new distinct values are increasingly rare, the DP prior implies n^* which is steadily increasing with n.

- The full prior for the number of distinct elements can also be derived:

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

  where the factors $c_n(m) = \Pr(n^* = m \mid \alpha = 1)$ can be computed using certain recurrence formulas (Antoniak, 1974; Escobar and West, 1995; Ghosal and van der Vaart, 2017).
  - If α has prior $p(\alpha)$, $\Pr(n^* = m) = \int \Pr(n^* = m \mid \alpha) p(\alpha) \mathrm{d}\alpha$.

Methods for posterior inference

- Data $= \{y_i, i = 1, \dots, n\}$ i.i.d., conditionally on G and ϕ, from $f(\cdot \mid G, \phi)$. (If the model includes a regression component, the data also include the covariate vectors x_i, and, in such cases, ϕ, typically, includes the vector of regression coefficients).

- Interest in inference for the latent mixing parameters $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$, for ϕ (and the hyperparameters α, ψ), for $f(y_0 \mid G, \phi)$, and, in general, for functionals $H(F(\cdot \mid G, \phi))$ of the random mixture $F(\cdot \mid G, \phi)$ (e.g., c.d.f. function, hazard function, mean and variance functionals, percentile functionals).

- Full inference, given the data, for all these random quantities is based on the joint posterior distribution of the DP mixture model

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

Marginal posterior simulation methods

- The joint posterior distribution can be expressed as

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

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

- $G \mid \boldsymbol{\theta}, \alpha, \psi \sim \text{DP}(\tilde{\alpha}, \tilde{G}_0)$, 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 c.d.f., $\tilde{G}_0(t) = \frac{\alpha}{\alpha+n} G_0(t \mid \psi) + \frac{1}{\alpha+n} \sum_{i=1}^{n} 1_{[\theta_i, \infty)}(t)$).

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

Marginal posterior simulation methods

- 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 and MacQueen, 1973),

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

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

$$p(\boldsymbol{\theta} \mid \alpha, \psi) = g_0(\theta_1 \mid \psi) \prod_{i=2}^{n} \left\{ \frac{\alpha}{\alpha + i - 1} g_0(\theta_i \mid \psi) + \frac{1}{\alpha + i - 1} \sum_{j=1}^{i-1} \delta_{\theta_j}(\theta_i) \right\}.$$

  Equivalently, the expression in terms of the DP induced partition structure can be used.

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

Marginal posterior simulation methods

- Therefore, the marginal posterior

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

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 $f(y_0 \mid G)$ (e.g., Lo, 1984; Brunner and Lo, 1989).
    - Approximations for special cases, e.g., for binomial DP mixtures (Berry and 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:

$$p(\theta_i \mid \{\theta_j : j \neq i\}, \alpha, \psi) = \frac{\alpha}{\alpha + n - 1} g_0(\theta_i \mid \psi) + \frac{1}{\alpha + n - 1} \sum_{j \neq i} \delta_{\theta_j}(\theta_i)$$

- **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 and West, 1995; Bush and MacEachern, 1996; Neal, 2000; Jain and Neal, 2004).

Simulation-based model fitting

- A key property for the implementation of the Gibbs sampler is the discreteness of G, which induces a partition (clustering) of the θ_i.
  - n^*: number of distinct elements (clusters) in the vector $(\theta_1, \dots, \theta_n)$.
  - θ_j^*, $j = 1, \dots, n^*$: the distinct θ_i.
  - $\boldsymbol{w} = (w_1, \dots, w_n)$: vector of configuration indicators, defined by $w_i = j$ if and only if $\theta_i = \theta_j^*$, $i = 1, \dots, n$.
  - n_j: size of j-th cluster, i.e., $n_j = |\{i : w_i = j\}|$, $j = 1, \dots, n^*$.
- $(n^*, \boldsymbol{w}, (\theta_1^*, \dots, \theta_{n^*}^*))$ is equivalent to $(\theta_1, \dots, \theta_n)$.
- Standard Gibbs sampler to draw from $p(\boldsymbol{\theta}, \phi, \alpha, \psi \mid \text{data})$ (Escobar and West, 1995) is based on the following full conditionals:
  1. $p(\theta_i \mid \{\theta_{i'} : i' \neq i\}, \alpha, \psi, \phi, \text{data})$, for $i = 1, \dots, n$.
  2. $p(\phi \mid \{\theta_i : i = 1, \dots, n\}, \text{data})$.
  3. $p(\psi \mid \{\theta_j^* : i = 1, \dots, n^*\}, n^*, \text{data})$.
  4. $p(\alpha \mid n^*, \text{data})$.

  (The expressions include conditioning only on the relevant variables, exploiting the conditional independence structure of the model and properties of the DP).

Simulation-based model fitting

1. For each $i = 1, \ldots, n$, $p(\theta_i \mid \{\theta_{i'} : i' \neq i\}, \alpha, \psi, \phi, \text{data})$ is simply a mixture of n^{*-} point masses and the posterior for θ_i based on y_i,

$$\frac{\alpha q_0}{\alpha q_0 + \sum_{j=1}^{n^{*-}} n_j^- q_j} h(\theta_i \mid \psi, \phi, y_i) + \sum_{j=1}^{n^{*-}} \frac{n_j^- q_j}{\alpha q_0 + \sum_{j=1}^{n^{*-}} n_j^- q_j} \delta_{\theta_j^{*-}}(\theta_i).$$

- $q_j = k(y_i \mid \theta_j^{*-}, \phi)$
- $q_0 = \int k(y_i \mid \theta, \phi) g_0(\theta \mid \psi) \mathrm{d}\theta$
- $h(\theta_i \mid \psi, \phi, y_i) \propto k(y_i \mid \theta_i, \phi) g_0(\theta_i \mid \psi)$
- g_0 is the density of G_0

- The superscript "$-$" denotes all relevant quantities when θ_i is removed from the vector $(\theta_1, \ldots, \theta_n)$, e.g., n^{*-} is the number of clusters in $\{\theta_{i'} : i' \neq i\}$.

- Updating θ_i implicitly updates w_i, $i = 1, \ldots, n$; before updating θ_{i+1}, we redefine n^*, θ_j^* for $j = 1, \ldots, n^*$, w_i for $i = 1, \ldots, n$, and n_j, for $j = 1, \ldots, n^*$.

Simulation-based model fitting

2. The posterior full conditional for ϕ does not involve the nonparametric part of the DP mixture model,

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

3. Regarding the parameters ψ of G_0,

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

   leading to standard updates under a conditionally conjugate prior $p(\psi)$.

Simulation-based model fitting

4. Although the posterior full conditional for α is not of a standard form, an augmentation method facilitates sampling if α has a gamma prior (say, with mean a_α / b_α) (Escobar and West, 1995),

$$\begin{aligned} p(\alpha \mid n^*, \text{data}) &\propto p(\alpha) \alpha^{n^*} \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)} \\ &\propto p(\alpha) \alpha^{n^* - 1} (\alpha + n) \text{Beta}(\alpha + 1, n) \\ &\propto p(\alpha) \alpha^{n^* - 1} (\alpha + n) \int_0^1 \eta^\alpha (1 - \eta)^{n-1} d\eta \end{aligned}$$

- Introduce an auxiliary variable η such that

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

- Extend the Gibbs sampler to draw $\eta \mid \alpha, \text{data} \sim \text{Beta}(\alpha + 1, n)$, and $\alpha \mid \eta, n^*, \text{data}$ from the two-component gamma mixture:

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

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

Improved marginal Gibbs sampler

- (West et al., 1994; Bush and MacEachern, 1996): adds one more step where the cluster locations θ_j^* are resampled at each iteration to improve the mixing of the chain.
- At each iteration, once step (1) is completed, we obtain a specific number of clusters n^* and configuration $\boldsymbol{w} = (w_1, \dots, w_n)$.
- After the marginalization over G, the prior for the θ_j^*, given the partition $(n^*, \boldsymbol{w})$, is given by $\prod_{j=1}^{n^*} g_0(\theta_j^* \mid \psi)$, i.e., given n^* and $\boldsymbol{w}$, the θ_j^* are i.i.d. from G_0.
- Hence, for each $j = 1,\dots,n^*$, the posterior full conditional

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

More general marginal MCMC algorithms

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

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

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

Posterior predictive distributions

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

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

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

- Or, equivalently, posterior samples

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

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

- Bayesian *density estimate* is based on the posterior predictive density $p(y_0 \mid \text{data})$ corresponding to a *new* y_0 (with associated θ_0).

Posterior predictive distributions

- Using, again, the Pólya urn structure for the DP,

$$p(\theta_0 \mid n^*, \mathrm{w}, \boldsymbol{\theta}^*, \alpha, \psi) = \frac{\alpha}{\alpha + n} g_0(\theta_0 \mid \psi) + \frac{1}{\alpha + n} \sum_{j=1}^{n^*} n_j \delta_{\theta_j^*}(\theta_0).$$

- The posterior predictive density is given by

$$p(y_0 \mid \text{data}) = \int \int k(y_0 \mid \theta_0, \phi) p(\theta_0 \mid n^*, \mathrm{w}, \boldsymbol{\theta}^*, \alpha, \psi) \\ p(n^*, \mathrm{w}, \boldsymbol{\theta}^*, \alpha, \psi, \phi \mid \text{data}) \mathrm{d}\theta_0 \mathrm{dwd}\boldsymbol{\theta}^* \mathrm{d}\alpha \mathrm{d}\psi \mathrm{d}\phi$$

- Hence, a sample $\{y_{0,b} : b = 1, \dots, B\}$ from the posterior predictive distribution can be obtained using the MCMC output, where, for each $b = 1, \dots, B$:
    - we first draw $\theta_{0,b}$ from $p(\theta_0 \mid n_b^*, \mathrm{w}_b, \boldsymbol{\theta}_b^*, \alpha_b, \psi_b)$
    - and then, draw $y_{0,b}$ from $K(\cdot \mid \theta_{0,b}, \phi_b)$.

Posterior predictive distributions

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

$$p(y_0 \mid \text{data}) =$$

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

- The integrand above is a mixture of:
  - the prior predictive density, $\mathrm{E}\{f(y_0 \mid G, \phi)\}$; and
  - a finite mixture with n^* components, with mixing parameters defined by the distinct θ_j^*, and weights given by n_j/n. This term dominates when α is small relative to n.

- The posterior predictive density for y_0 is obtained by averaging this mixture with respect to the posterior distribution 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 $f(y_0 \mid G, \phi)$ (at point y_0), i.e., $p(y_0 \mid \text{data}) = \mathrm{E}(f(y_0 \mid G, \phi) \mid \text{data})$.
    - The Bayesian density estimate under a DP mixture model can be obtained without sampling from the posterior distribution of G.

- Analogously, we can obtain posterior moments for $H(F(\cdot \mid G, \phi)) = \int H(K(\cdot \mid \theta, \phi))\mathrm{d}G(\theta)$, where H is a linear functional (Gelfand and 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 c.d.f. functionals, mean functional).

- How about more general types of inference?
    - Interval estimates for $F(y_0 \mid G, \phi)$ or $f(y_0 \mid G, \phi)$, for specified y_0?
    - Inference for non-linear functions of the c.d.f., e.g., cumulative hazard, $-\log(1 - F(y_0 \mid G, \phi))$, or hazard, $f(y_0 \mid G, \phi)/(1 - F(y_0 \mid G, \phi))$, functions?
    - Inference for other non-linear functionals, e.g., for percentiles?

Inference for general functionals of the random mixture

- Such inferences require the posterior distribution of G. Recall

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

  and

$$G \mid \boldsymbol{\theta}, \alpha, \psi \sim \text{DP}\left(\alpha + n, \tilde{G}_0(\cdot) = \frac{\alpha}{\alpha + n} G_0(\cdot \mid \psi) + \frac{1}{\alpha + n} \sum\nolimits_{i=1}^{n} \delta_{\theta_i}(\cdot)\right)$$

- Hence, given posterior samples $(\boldsymbol{\theta}_b, \alpha_b, \psi_b, \phi_b)$, for $b = 1, \ldots, B$, from the marginalized version of the DP mixture, we can draw G_b from $p(G \mid \boldsymbol{\theta}_b, \alpha_b, \psi_b)$ using:
  - The original DP definition if we only need sample paths for the c.d.f. of the mixture (and y is univariate) (e.g., Krnjajić et al., 2008).
  - More generally, the DP constructive definition with a truncation approximation (Gelfand and Kottas, 2002; Ishwaran and Zarepour, 2002).

Inference for general functionals of the random mixture

- Applying directly the DP constructive definition,

$$G_b = \zeta_1 \, \delta_{U_1} + \sum_{\ell=2}^{L-1} \left\{ \zeta_\ell \prod_{r=1}^{\ell-1} (1-\zeta_r) \right\} \delta_{U_\ell} + \left\{ \prod_{r=1}^{L-1} (1-\zeta_r) \right\} \delta_{U_L}$$

  where the ζ_ℓ, $\ell = 1, ..., L-1$, are i.i.d. Beta(1,$\alpha + n$), and (independently) the U_ℓ, $\ell = 1, ..., L$, are i.i.d. $\tilde{G}_0$.

- A more efficient truncation approximation through an alternative representation for the conditional posterior of G (Pitman, 1996)

$$G \mid (n^*, \boldsymbol{w}, \boldsymbol{\theta}^*), \alpha, \psi \overset{\mathcal{D}}{=} q_{n^*+1} G^* + \sum_{j=1}^{n^*} q_j \, \delta_{\theta_j^*}$$

  where $G^* \mid \alpha, \psi \sim \mathrm{DP}(\alpha, G_0(\psi))$ and, independently of G^*, the vector of weights, $(q_1, ..., q_{n^*}, q_{n^*+1}) \mid \alpha, \boldsymbol{w} \sim \mathrm{Dirichlet}(n_1, ..., n_{n^*}, \alpha)$.

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

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:

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

  where $G_0(\mu, \sigma^2) = \mathrm{N}(\mu \mid \mu_0, \sigma^2/\kappa)\mathrm{IGamma}(\sigma^2 \mid \nu, s)$.

- We consider four different prior specifications to explore the effect of increasing flexibility in the DP prior hyperparameters.
- Figure 3.2 shows posterior predictive density estimates obtained using the function `DPdensity` in the `R` package `DPpackage`.

Density estimation data example

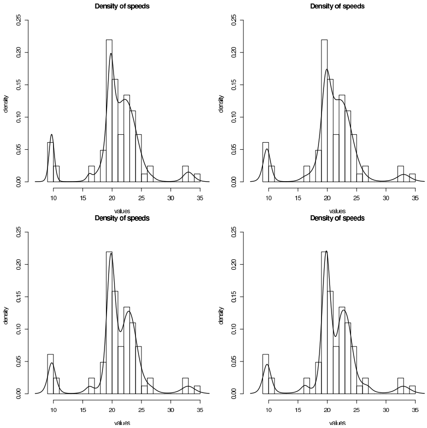


Figure 3.2: Histograms of the raw data and posterior predictive densities under four prior choices for the `galaxy` data. In the top left panel we set $\alpha = 1$, $\mu_0 = 0$, $s = 2$, $\nu = 4$, $\kappa \sim \text{Gam}(0.5, 50)$; the top right panel uses the same settings except $s \sim \text{IGamma}(4, 2)$; 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 $\alpha \sim \text{Gam}(2, 2)$.

Conditional posterior simulation methods

- The main characteristic of the marginal MCMC methods is that they are based on the posterior distribution 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* or *collapsed* methods).

- Although posterior inference for G is possible under the collapsed sampler, it is of interest to study alternative *conditional* posterior simulation approaches that impute G as part of the MCMC algorithm, and also improve on the mixing of marginal samplers.
  - Methods based on finite truncation approximation of G, using its stick-breaking representation – main example: Blocked Gibbs sampler (Ishwaran and Zarepour, 2000; Ishwaran and James, 2001).
  - Other approaches based on retrospective sampling techniques (Papaspiliopoulos and Roberts, 2008), slice sampling methods (Walker, 2007; Kalli et al., 2011), as well as combinations of retrospective and slice sampling (Yau et al., 2011).

Blocked Gibbs sampler

- Builds from truncation approximation to mixing distribution G given, for finite N, by

$$G_N = \sum_{\ell=1}^{N} p_\ell \, \delta_{Z_\ell}$$

  - The Z_ℓ, $\ell = 1, \dots, N$, are i.i.d. G_0.
  - The weights arise through stick-breaking (with truncation)

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

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

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

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

The generalized Dirichlet distribution

- Assume that $V_\ell \overset{ind.}{\sim} \text{Beta}(a_\ell, b_\ell)$, for $\ell = 1, ..., N-1$, and define a probability vector, $\boldsymbol{p} = (p_1, ..., p_N)$, through

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

- Then, $\boldsymbol{p}$ follows a generalized Dirichlet distribution, with parameters $\boldsymbol{a} = (a_1, ..., a_{N-1})$ and $\boldsymbol{b} = (b_1, ..., b_{N-1})$, and with density given by

$$f(\boldsymbol{p} \mid \boldsymbol{a}, \boldsymbol{b}) = \left\{ \prod_{\ell=1}^{N-1} \frac{\Gamma(a_\ell + b_\ell)}{\Gamma(a_\ell)\Gamma(b_\ell)} \right\} p_1^{a_1 - 1} \times \ldots \times p_{N-1}^{a_{N-1}-1} p_N^{b_{N-1}-1} (1 - p_1)^{b_1 - (a_2 + b_2)}$$

$$(1 - (p_1 + p_2))^{b_2 - (a_3 + b_3)} \times \ldots \times \left(1 - \sum\nolimits_{\ell=1}^{N-2} p_\ell\right)^{b_{N-2} - (a_{N-1} + b_{N-1})}$$

- If $b_{\ell-1} = a_\ell + b_\ell$, for $\ell = 2, ..., N-1$, the distribution reduces to a Dirichlet$(c_1, ..., c_N)$ with $c_\ell = a_\ell$, for $\ell = 1, ..., N-1$, and $c_N = b_{N-1}$.

Truncation level specification

- The DP truncation level N can be chosen to any desired level of accuracy.

- A simple approach based on the prior expectation for the partial sum of DP stick-breaking weights, $\mathrm{E}(\sum_{\ell=1}^{N} \omega_\ell \mid \alpha) = 1 - \{\alpha/(\alpha+1)\}^{N}$ (can be averaged over the prior for α to estimate $\mathrm{E}(\sum_{\ell=1}^{N} \omega_\ell)$).
  - For example, $\mathrm{E}(\sum_{\ell=1}^{25} \omega_\ell \mid \alpha = 2) = 0.99996$, and $\mathrm{E}(\sum_{\ell=1}^{75} \omega_\ell) = 0.99997$ under an exponential prior for α with mean 2.

- A more general approach, which involves also the sample size n, is available through Th. 2 in Ishwaran and James (2001): approximate upper bound of $4n\exp\{-(N-1)/\alpha\}$ on the L_1 distance between the prior predictive probability of the sample under the countable representation for G and its truncated version G_N.
  - For example, with $\alpha = 2$, the bound is 0.00001656 for $n = 10^2$ and $N = 35$, and it is 0.00001678 for $n = 10^7$ and $N = 58$.

Blocked Gibbs sampler

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

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

- 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,

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

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

Blocked Gibbs sampler

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

- Final version of the hierarchical model:

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

- Marginalizing over the L_i, we obtain the same finite mixture model for the y_i: $f(y \mid \boldsymbol{p}, \mathrm{Z}, \phi) = \sum_{\ell=1}^{N} p_\ell \, k(y \mid Z_\ell, \phi)$.

Posterior full conditional distributions

1. To update Z_ℓ for $\ell = 1, \dots, N$:
   - Let n^* be the number of distinct values $\{L_j^* : j = 1, \dots, n^*\}$ of vector L.
   - Then, the posterior full conditional for Z_ℓ, $\ell = 1, \dots, N$, can be expressed in general as:

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

   - If $\ell \notin \{L_j^* : j = 1, \dots, n^*\}$, Z_ℓ is drawn from $G_0(\cdot \mid \psi)$
   - For $\ell = L_j^*$, $j = 1, \dots, n^*$,

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

2. The posterior full conditional for $\boldsymbol{p}$ is

$$p(\boldsymbol{p} \mid \ldots, \text{data}) \propto f(\boldsymbol{p} \mid \alpha) \prod_{\ell=1}^{N} p_{\ell}^{M_{\ell}},$$

   where $M_{\ell} = |\{i : L_i = \ell\}|$, $\ell = 1, \ldots, N$.

   - Results in a generalized Dirichlet distribution, which can be sampled through independent latent Beta variables.
   - $V_{\ell}^{*} \overset{ind.}{\sim} \text{Beta}(1 + M_{\ell}, \alpha + \sum_{r=\ell+1}^{N} M_r)$, for $\ell = 1, \ldots, N-1$.
   - $p_1 = V_1^{*}$; $p_{\ell} = V_{\ell}^{*} \prod_{r=1}^{\ell-1}(1 - V_r^{*})$, for $\ell = 2, \ldots, N-1$; and $p_N = 1 - \sum_{\ell=1}^{N-1} p_{\ell}$.

3. Updating the L_i, $i = 1, \ldots, n$:
   - Each L_i is drawn from the discrete distribution on $\{1, \ldots, N\}$ with probabilities $\tilde{p}_{\ell i} \propto p_{\ell} k(y_i \mid Z_{\ell}, \phi)$, for $\ell = 1, \ldots, N$.
   - Note that 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.

4. The posterior full conditional for ϕ is

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

5. The posterior full conditional for ψ is

$$p(\psi \mid \ldots, \text{data}) \propto p(\psi) \prod_{j=1}^{n^*} g_0(Z_{L_j^*} \mid \psi).$$

6. The posterior full conditional for α is proportional to $p(\alpha)\alpha^{N-1}p_N^{\alpha}$, which with a gamma$(a_\alpha, b_\alpha)$ prior for α, results in a gamma$(N+a_\alpha-1, b_\alpha-\log(p_N))$ distribution. (For numerical stability, compute $\log(p_N) = \log\prod_{r=1}^{N-1}(1 - V_r^*) = \sum_{r=1}^{N-1}\log(1 - V_r^*)$.)

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

Posterior predictive inference

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

$$p(y_0 \mid \text{data}) = \int k(y_0 \mid Z_{L_0}, \phi) \left(\sum_{\ell=1}^{N} p_\ell \delta_\ell(L_0) \right)$$
$$p(\mathrm{Z}, \mathrm{p}, \mathrm{L}, \phi, \alpha, \psi \mid \text{data}) \mathrm{d}L_0 \, \mathrm{dZ} \, \mathrm{dL} \, \mathrm{d}\boldsymbol{p} \, \mathrm{d}\phi \, \mathrm{d}\alpha \, \mathrm{d}\psi$$
$$= \int \left(\sum_{\ell=1}^{N} p_\ell k(y_0 \mid Z_\ell, \phi) \right)$$
$$p(\mathrm{Z}, \boldsymbol{p}, \mathrm{L}, \phi, \alpha, \psi \mid \text{data}) \mathrm{dZ} \, \mathrm{dL} \, \mathrm{d}\boldsymbol{p} \, \mathrm{d}\phi \, \mathrm{d}\alpha \, \mathrm{d}\psi$$
$$= \mathrm{E}(f(y_0 \mid \boldsymbol{p}, \mathrm{Z}, \phi) \mid \text{data}).$$

- Hence, $p(y_0 \mid \text{data})$ can be estimated over a grid in y_0 by drawing samples $\{L_{0b} : b = 1, \dots, B\}$ for L_0, based on the posterior samples for p, and computing the Monte Carlo estimate

$$B^{-1} \sum\nolimits_{b=1}^{B} k(y_0 \mid Z_{L_{0b}}, \phi_b),$$

where B is the posterior sample size.

Model checking/comparison for DP mixtures

- Posterior predictive estimation/sampling is straightforward for DP mixtures, and this allows using standard model checking/comparison techniques for (hierarchical) Bayesian models. Two examples are discussed next.

- Posterior predictive loss criterion (Gelfand and Ghosh, 1998): choose model that minimizes $D_k(M) = P(M) + \{k/(k+1)\}G(M)$, where:
    - $P(M) = \sum_{i=1}^{n} \mathrm{Var}^{(M)}(y_{new,i} \mid \text{data})$ is a penalty term, and
    - $G(M) = \sum_{i=1}^{n} \{y_i - \mathrm{E}^{(m)}(y_{new,i} \mid \text{data})\}^2$ is a goodness of fit term.
    - $\mathrm{E}^{(M)}(y_{new,i} \mid \text{data})$ and $\mathrm{Var}^{(M)}(y_{new,i} \mid \text{data})$ is the posterior predictive mean and posterior predictive variance under model M for replicated response $y_{new,i}$; in regression problems, the posterior predictive distribution for $y_{new,i}$ is evaluated for the observed vector of covariates $\boldsymbol{x}_i$.
    - $k \geq 0$ controls the weight assigned to the goodness of fit term.

Model checking/comparison for DP mixtures

- Conditional predictive ordinate (CPO) for observation y_i under model M: $\text{CPO}_i^{(M)} = p^{(M)}(y_i \mid \{y_j : j \neq i\})$, that is, the value of the posterior predictive density at y_i, given the data set excluding y_i.
  - Ratio $\text{CPO}_i^{(M_1)}/\text{CPO}_i^{(M_2)}$ describes how well model M_1 supports observation y_i relative to model M_2.
  - "Pseudo Bayes factor", $B_{12} = \prod_{i=1}^{n}(\text{CPO}_i^{(M_1)}/\text{CPO}_i^{(M_2)})$, is an aggregate summary of how well supported the data are by model M_1 relative to model M_2 (Geisser and Eddy, 1979).
  - "Log pseudo marginal likelihood" (LPML) for model M: $\text{LPML}_M = \log \prod_{i=1}^{n} \text{CPO}_i^{(M)}$, such that $B_{12} = \exp(\text{LPML}_{M_1} - \text{LPML}_{M_2})$.

- The Bayes factor requires the non-trivial computation of the DP mixture model marginal likelihood, $m(\boldsymbol{y})$, where $\boldsymbol{y} = (y_1, ..., y_n)$.
  - $m(\boldsymbol{y}) = \int L(\boldsymbol{y}; \phi, \alpha, \psi) p(\phi) p(\alpha) p(\psi) \, \text{d}\phi \text{d}\alpha \text{d}\psi$
  - $L(\boldsymbol{y}; \phi, \alpha, \psi) = \int \{\prod_{i=1}^{n} k(y_i \mid \theta_i, \phi)\} p(\boldsymbol{\theta} \mid \alpha, \psi) \, \text{d}\boldsymbol{\theta}$
  - One approach is given in Basu and Chib (2003), using sequential importance sampling to estimate the likelihood ordinate $L(\boldsymbol{y}; \phi, \alpha, \psi)$.

Alternative computational inference schemes

- Alternative (to MCMC) fitting techniques have been studied.
    - Sequential importance sampling (Liu, 1996; Quintana, 1998; MacEachern et al., 1999; Quintana and Newton, 2000; Carvalho et al., 2010).
    - Weighted Chinese restaurant algorithms (Ishwaran and Takahara, 2002; Ishwaran and James 2003).
    - Monte Carlo EM (Naskar and Das, 2004).
    - Predictive recursion (Newton and Zhang, 1999; Tokdar et al., 2009).
    - Variational algorithms (e.g., Blei and Jordan, 2006; Zobay, 2009).

- Posterior simulation for DP mixture models (and, more generally, Bayesian nonparametric models) for *large* datasets is an active area of research – some of the earlier contributions to scalable NPB methods include Guha (2010) and Wang and Dunson (2011).

Variational algorithms

- An alternative to MCMC methods which is very popular in the machine learning literature, and is gaining some traction within the statistics community; see Blei et al. (2017) for a review.

- Consider a generic model where $\boldsymbol{y}$ denotes the data and $\boldsymbol{\theta}$ collects all parameters. Variational algorithms aim at replacing the intractable posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ with a more tractable approximation $q_{\boldsymbol{\eta}}(\boldsymbol{\theta})$ whose parameters $\boldsymbol{\eta}$ are chosen to minimize

$$K(p||q) = \int \log\left(\frac{q_{\boldsymbol{\eta}}(\boldsymbol{\theta})}{p(\boldsymbol{\theta} \mid \boldsymbol{y})}\right) q_{\boldsymbol{\eta}}(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}$$

  the Kullback-Leibler divergence between $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ and $q_{\boldsymbol{\eta}}(\boldsymbol{\theta})$.

- Variational inference methods reformulate the problem of computing the posterior distribution as an easier (and faster!) to handle optimization problem. The main drawback is that, in contrast to MCMC methods, there is no general theory that ensures convergence to the posterior distribution.

Variational algorithms

- The minimization of $K(p||q)$ can be alternatively approached as maximization of a lower bound on the log marginal likelihood:

$$\log p(\boldsymbol{y}) \geq \mathrm{E}_q \left\{ \log p(\boldsymbol{\theta}, \boldsymbol{y}) \right\} - \mathrm{E}_q \left\{ \log q_{\boldsymbol{\eta}}(\boldsymbol{\theta}) \right\}$$

  - The gap in the bound is the K-L divergence between $q_{\boldsymbol{\eta}}$ and the true posterior: $\log p(\boldsymbol{y}) = \mathrm{E}_q \left\{ \log p(\boldsymbol{\theta}, \boldsymbol{y}) \right\} - \mathrm{E}_q \left\{ \log q_{\boldsymbol{\eta}}(\boldsymbol{\theta}) \right\} + K(p||q)$
  - Recall that $K(p||q) \geq 0$ (with equality if-f $p = q$).

- Two key ingredients for an efficient algorithm: the particular form of the variational distribution $q_{\boldsymbol{\eta}}$, and the optimization procedure.

- In principle, there is a lot of freedom in choosing $q_{\boldsymbol{\eta}}$. Practical limitations arise from the need to have a tractable approximation and to compute the expectations $\mathrm{E}_q \left\{ \log q_{\boldsymbol{\eta}}(\boldsymbol{\theta}) \right\}$ and $\mathrm{E}_q \left\{ \log p(\boldsymbol{\theta}, \boldsymbol{y}) \right\}$.

Mean-field variational algorithms

- *Mean-field* methods use a factorized variational distribution

$$q_{\boldsymbol{\eta}}(\boldsymbol{\theta}) = \prod_{k=1}^{K} q_{k,\boldsymbol{\eta}_k}(\theta_k)$$

  where $\boldsymbol{\theta} = (\theta_1, ..., \theta_K)$, and are very popular due to their tractability.

- In particular, if the conditional posterior distribution for each θ_k belongs to the exponential family, it is natural to select $q_{k,\boldsymbol{\eta}_k}(\theta_k)$ as a member of the conditionally conjugate prior family.
  - Then, the optimization of the K-L divergence w.r.t. a single variational parameter $\boldsymbol{\eta}_k$ is achieved by computing the expectation (w.r.t. $q_{\boldsymbol{\eta}}$) of the exponential family natural parameter for θ_k.
  - Recursively updating each $\boldsymbol{\eta}_k$ by computing this expectation corresponds to performing coordinate ascent in the K-L divergence.
  - Hence, the algorithm has the flavor of a Gibbs sampler.

Mean-field variational algorithms

- More specifically:
  - assume the posterior full conditional distribution for θ_k is a member of the exponential family

    $$p(\theta_k \mid \boldsymbol{\theta}_{(-k)}, \boldsymbol{y}) \propto \exp\left\{ \sum_{r=1}^{p} A_{k,r}(\boldsymbol{\theta}_{(-k)}, \boldsymbol{y})\, B_r(\theta_k) \;-\; C(\theta_k) \right\}$$

    where $\boldsymbol{\theta}_{(-k)} = (\theta_1, ..., \theta_{k-1}, \theta_{k+1}, ..., \theta_K)$
  - take the variational approximation to be

    $$q_{k,\boldsymbol{\eta}_k}(\theta_k) \propto \exp\left\{ \sum_{r=1}^{p} \eta_{k,r}\, B_r(\theta_k) \;-\; H(\theta_k) \right\}$$

    where $\boldsymbol{\eta}_k = (\eta_{k,1}, ..., \eta_{k,p})$
  - then, the maximum of the log marginal likelihood lower bound w.r.t. $\eta_{k,r}$ (fixing all other variational parameters) is attained at

    $$\hat{\eta}_{k,r} = \mathrm{E}_q \left\{ A_{k,r}(\boldsymbol{\theta}_{(-k)}, \boldsymbol{y}) \right\}$$

(see, e.g., Appendix A in Blei and Jordan, 2006)

Mean-field variational algorithms

- Coordinate ascent algorithm: iteratively maximize the lower bound w.r.t. each $\boldsymbol{\eta}_k$ holding the other variational parameters fixed.
  - For models with posterior full conditionals in the exponential family, each iteration of the algorithm involves recursive computing of expectations (much less computationally intensive than simulation).

- Variational approximations are relatively straightforward to work with in conditionally conjugate models. Extensions for non-conjugate models are more challenging.

- In general, (mean-field) variational algorithms are sensitive to initial values (under conditions, the coordinate ascent algorithm finds a local maximum), tend to underestimate the uncertainty in the posterior distribution, and can not capture the dependence among parameters.

- Main inference focus on posterior predictive estimation; inference for more general functionals becomes more computationally intensive.

- An early example of mean-field methods for DP mixture models with exponential family kernels and the corresponding conjugate prior as the DP centering distribution (Blei and Jordan, 2006).

Mean-field variational methods for DP mixtures

- **Example:** location normal mixture model

- Consider the (blocked Gibbs sampler) truncated DP mixture model formulation (with fixed DP prior hyperparameters)

$$\begin{aligned}
y_i \mid \boldsymbol{Z}, L_i, \phi &\overset{ind.}{\sim} \mathrm{N}(y_i \mid Z_{L_i}, \phi^{-1}) && i = 1, \dots, n \\
L_i \mid \boldsymbol{V} &\overset{i.i.d.}{\sim} \prod_{\ell=1}^{N} (p_\ell(\boldsymbol{V}))^{1(L_i = \ell)} && i = 1, \dots, n \\
Z_\ell &\overset{i.i.d.}{\sim} \mathrm{N}(Z_\ell \mid m, s^2) && \ell = 1, \dots, N \\
V_\ell &\overset{i.i.d.}{\sim} \mathrm{Beta}(V_\ell \mid 1, \alpha) && \ell = 1, \dots, N-1 \\
\phi &\sim \mathrm{ga}(\phi \mid a_\phi, b_\phi)
\end{aligned}$$

where $\mathrm{ga}(a, b)$ is the gamma distribution with mean a/b, and

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

Mean-field variational methods for DP mixtures

- Parameter vector, $\boldsymbol{\theta} = (\boldsymbol{Z}, \boldsymbol{V}, \boldsymbol{L}, \phi)$, with $p(\boldsymbol{\theta}, \boldsymbol{y})$ given by

$$\text{ga}(\phi \mid a_\phi, b_\phi) \prod_{\ell=1}^{N-1} \text{Beta}(V_\ell \mid 1, \alpha) \prod_{\ell=1}^{N} \text{N}(Z_\ell \mid m, s^2) \prod_{i=1}^{n} p(L_i \mid \boldsymbol{V}) \prod_{i=1}^{n} \text{N}(y_i \mid Z_{L_i}, \phi^{-1})$$

where $p(L_i \mid \boldsymbol{V}) = \prod_{\ell=1}^{N} (p_\ell(\boldsymbol{V}))^{1(L_i=\ell)} = \prod_{\ell=1}^{N-1} V_\ell^{1(L_i=\ell)} (1 - V_\ell)^{1(L_i>\ell)}$

- Mean-field variational approximation:

$$q_{\boldsymbol{\eta}}(\boldsymbol{\theta}) = q_{\boldsymbol{\beta}}(\phi) \prod_{\ell=1}^{N-1} q_{\boldsymbol{\gamma}_\ell}(V_\ell) \prod_{\ell=1}^{N} q_{\boldsymbol{\xi}_\ell}(Z_\ell) \prod_{i=1}^{n} q_{\boldsymbol{\pi}_i}(L_i)$$

where $q_{\boldsymbol{\beta}}(\phi) = \text{ga}(\phi \mid \beta_1, \beta_2)$, $q_{\boldsymbol{\gamma}_\ell}(V_\ell) = \text{Beta}(V_\ell \mid \gamma_{\ell 1}, \gamma_{\ell 2})$, $q_{\boldsymbol{\xi}_\ell}(Z_\ell) = \text{N}(Z_\ell \mid \xi_{\ell 1}, \xi_{\ell 2})$, and $q_{\boldsymbol{\pi}_i}(L_i) = \prod_{\ell=1}^{N} \pi_{i\ell}^{1(L_i=\ell)}$

- Full set of variational parameters

$$\boldsymbol{\eta} = (\beta_1, \beta_2, \{(\gamma_{\ell 1}, \gamma_{\ell 2})\}_{\ell=1}^{N-1}, \{(\xi_{\ell 1}, \xi_{\ell 2})\}_{\ell=1}^{N}, \{(\pi_{i1}, \dots, \pi_{iN})\}_{i=1}^{n})$$

Mean-field variational methods for DP mixtures

- Updating parameters $\{(\gamma_{\ell 1}, \gamma_{\ell 2}) : \ell = 1, ..., N-1\}$
- For any $\ell = 1, ..., N-1$

$$p(V_\ell \mid \ldots, \boldsymbol{y}) = \text{Beta}\left(1 + \sum_{i=1}^{n} 1(L_i = \ell), \alpha + \sum_{i=1}^{n} 1(L_i > \ell)\right)$$
$$\propto \exp\left\{\left(\sum_{i=1}^{n} 1(L_i = \ell)\right)\log(V_\ell) + \left(\alpha - 1 + \sum_{i=1}^{n} 1(L_i > \ell))\right)\log(1 - V_\ell)\right\}$$

- Variational approximation

$$q_{\boldsymbol{\gamma}_\ell}(V_\ell) = \text{Beta}(V_\ell \mid \gamma_{\ell 1}, \gamma_{\ell 2}) \propto \exp\{(\gamma_{\ell 1} - 1)\log(V_\ell) + (\gamma_{\ell 2} - 1)\log(1 - V_\ell)\}$$

- Therefore, $\hat{\gamma}_{\ell 1} - 1 = \text{E}_q(\sum_{i=1}^{n} 1(L_i = \ell))$, and $\hat{\gamma}_{\ell 2} - 1 = \text{E}_q(\alpha - 1 + \sum_{i=1}^{n} 1(L_i > \ell))$, which yields

$$\hat{\gamma}_{\ell 1} = 1 + \sum_{i=1}^{n} \pi_{i\ell} \quad \text{and} \quad \hat{\gamma}_{\ell 2} = \alpha + \sum_{i=1}^{n} \sum_{r=\ell+1}^{N} \pi_{ir}$$

Mean-field variational methods for DP mixtures

- Updating parameters $\{(\xi_{\ell 1}, \xi_{\ell 2}) : \ell = 1, ..., N\}$
- For any $\ell = 1, ..., N$, the posterior full conditional for Z_ℓ is normal with mean $= (m + s^2 \phi \sum_{i=1}^n y_i 1(L_i = \ell))/(1 + s^2 \phi \sum_{i=1}^n 1(L_i = \ell))$, and variance $= s^2/(1 + s^2 \phi \sum_{i=1}^n 1(L_i = \ell))$. Therefore,

$$p(Z_\ell \mid ..., \mathbf{y}) \propto \exp\left\{ \left(\frac{m + s^2 \phi \sum_{i=1}^n y_i 1(L_i = \ell))}{s^2} \right) Z_\ell - \left(\frac{1 + s^2 \phi \sum_{i=1}^n 1(L_i = \ell)}{2s^2} \right) Z_\ell^2 \right\}$$

- Variational approximation

$$q_{\boldsymbol{\xi}_\ell}(Z_\ell) = \mathrm{N}(Z_\ell \mid \xi_{\ell 1}, \xi_{\ell 2}) \propto \exp\left\{ (\xi_{\ell 1}/\xi_{\ell 2}) Z_\ell - (1/2\xi_{\ell 2}) Z_\ell^2 \right\}$$

- Therefore, $\hat{\xi}_{\ell 2}^{-1} = s^{-2}\{1 + s^2 \mathrm{E}_q(\phi)\, \mathrm{E}_q(\sum_{i=1}^n 1(L_i = \ell))\}$, and $\hat{\xi}_{\ell 1}\hat{\xi}_{\ell 2}^{-1} = s^{-2}\{m + s^2 \mathrm{E}_q(\phi)\, \mathrm{E}_q(\sum_{i=1}^n y_i 1(L_i = \ell))\}$, which yields

$$\hat{\xi}_{\ell 1} = \frac{m + s^2 \beta_1 \beta_2^{-1} \sum_{i=1}^n y_i \, \pi_{i\ell}}{1 + s^2 \beta_1 \beta_2^{-1} \sum_{i=1}^n \pi_{i\ell}} \quad \text{and} \quad \hat{\xi}_{\ell 2} = \left(s^{-2} + \beta_1 \beta_2^{-1} \sum_{i=1}^n \pi_{i\ell} \right)^{-1}$$

Mean-field variational methods for DP mixtures

- Updating parameters $\{(\pi_{i1}, \ldots, \pi_{iN}) : i = 1, \ldots, n\}$
- For $i = 1, \ldots, n$, $\Pr(L_i = \ell \mid \ldots, \boldsymbol{y}) \propto p_\ell(\boldsymbol{V}) \, \mathrm{N}(y_i \mid Z_\ell, \phi^{-1})$, for $\ell = 1, \ldots, N$, so

$$p(L_i \mid \ldots, \boldsymbol{y}) \propto \exp\left\{ \sum_{\ell=1}^{N} 1(L_i = \ell) \, \log\{\phi^{1/2} \, \exp(-0.5 \, \phi \, (y_i - Z_\ell)^2) \, p_\ell(\boldsymbol{V})\} \right\}$$

- Variational approximation: $q_{\boldsymbol{\pi}_i}(L_i) = \prod_{\ell=1}^{N} \pi_{i\ell}^{1(L_i=\ell)} \propto \exp\left\{ \sum_{\ell=1}^{N} 1(L_i = \ell) \, \log(\pi_{i\ell}) \right\}$
- So, for $\ell = 2, \ldots, N-1$ (the expressions for $\ell = 1$ and $\ell = N$ are special cases):

$$\log(\hat{\pi}_{i\ell}) \propto 0.5 \, \mathrm{E}_q(\log(\phi)) - 0.5 \, \mathrm{E}_q(\phi) \, \mathrm{E}_q\{(y_i - Z_\ell)^2\} + \mathrm{E}_q(\log(V_\ell)) + \sum\nolimits_{r=1}^{\ell-1} \mathrm{E}_q(\log(1 - V_r))$$

which yields $\hat{\pi}_{i\ell} \propto \exp(W_\ell)$, where

$$W_\ell = 0.5\{\Psi(\beta_1) - \log(\beta_2)\} - 0.5\beta_1\beta_2^{-1}\{y_i^2 - 2y_i\xi_{\ell 1} + \xi_{\ell 2} + \xi_{\ell 1}^2\}$$
$$+ \{\Psi(\gamma_{\ell 1}) - \Psi(\gamma_{\ell 1} + \gamma_{\ell 2})\} + \sum\nolimits_{r=1}^{\ell-1} \{\Psi(\gamma_{r2}) - \Psi(\gamma_{r1} + \gamma_{r2})\}$$

using two results for the gamma and Beta distributions:

- if $X \sim \mathrm{ga}(\alpha, \beta)$, with mean α/β, then $\mathrm{E}(\log(X)) = \Psi(\alpha) - \log(\beta)$, where $\Psi(\alpha) = \frac{d}{d\alpha} \log(\Gamma(\alpha)) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$ is the digamma function.
- if $X \sim \mathrm{Beta}(\alpha, \beta)$, with mean $\alpha/(\alpha + \beta)$, then $\mathrm{E}(\log(X)) = \Psi(\alpha) - \Psi(\alpha + \beta)$.

Mean-field variational methods for DP mixtures

- Updating parameters (β_1, β_2)
- $p(\phi \mid ..., \boldsymbol{y})$ is a gamma distribution with shape and rate parameter given by $a_\phi + 0.5 \sum_{\ell=1}^{N} \sum_{i=1}^{n} 1(L_i = \ell)$ and $b_\phi + 0.5 \sum_{\ell=1}^{N} \sum_{i=1}^{n} 1(L_i = \ell)(y_i - Z_\ell)^2$, respectively. Therefore, $p(\phi \mid ..., \boldsymbol{y})$ can be written proportional to

$$\exp\left\{ \left(a_\phi - 1 + 0.5 \sum_{\ell=1}^{N} \sum_{i=1}^{n} 1(L_i = \ell) \right) \log(\phi) - \left(b_\phi + 0.5 \sum_{\ell=1}^{N} \sum_{i=1}^{n} 1(L_i = \ell)(y_i - Z_\ell)^2 \right) \phi \right\}$$

- Variational approximation

$$q_{\boldsymbol{\beta}}(\phi) = \text{ga}(\phi \mid \beta_1, \beta_2) \propto \exp\{(\beta_1 - 1)\log(\phi) - \beta_2\,\phi\}$$

- Therefore, $\hat{\beta}_1 = a_\phi + 0.5 \sum_{i=1}^{n} \sum_{\ell=1}^{N} \text{E}_q\{1(L_i = \ell)\}$, and $\hat{\beta}_2 = b_\phi + 0.5 \sum_{\ell=1}^{N} \sum_{i=1}^{n} \text{E}_q\{1(L_i = \ell)\}\, \text{E}_q\{(y_i - Z_\ell)^2\}$, which yields

$$\hat{\beta}_1 = a_\phi + 0.5\, n \quad \text{and} \quad \hat{\beta}_2 = b_\phi + 0.5 \sum_{\ell=1}^{N} \sum_{i=1}^{n} \pi_{i\ell}\,(y_i^2 - 2y_i\xi_{\ell 1} + \xi_{\ell 2} + \xi_{\ell 1}^2)$$

Mean-field variational methods for DP mixtures

- Approximation to the posterior predictive density, replacing the posterior distribution with the (estimated) variational distribution:

$$
\begin{aligned}
p(y_0 \mid \boldsymbol{y}) &= \int \left\{ \sum_{\ell=1}^{N} p_\ell(\boldsymbol{V})\, \mathrm{N}(y_0 \mid Z_\ell, \phi^{-1}) \right\} p(\boldsymbol{\theta} \mid \boldsymbol{y}) \mathrm{d}\boldsymbol{\theta} \\
&\approx \int \left\{ \sum_{\ell=1}^{N} p_\ell(\boldsymbol{V})\, \mathrm{N}(y_0 \mid Z_\ell, \phi^{-1}) \right\} q_{\hat{\boldsymbol{\beta}}}(\phi) \prod_{\ell=1}^{N-1} q_{\hat{\gamma}_\ell}(V_\ell) \prod_{\ell=1}^{N} q_{\hat{\boldsymbol{\xi}}_\ell}(Z_\ell)\, \mathrm{d}\phi \mathrm{d}\boldsymbol{V} \mathrm{d}\boldsymbol{Z} \\
&\approx \sum_{\ell=1}^{N} \mathrm{E}_q\{p_\ell(\boldsymbol{V})\}\, \mathrm{E}_q\{\mathrm{N}(y_0 \mid Z_\ell, \phi^{-1})\}
\end{aligned}
$$

- $\mathrm{E}_q\{p_\ell(\boldsymbol{V})\} = \mathrm{E}_q\{V_\ell \prod_{r=1}^{\ell-1}(1-V_r)\} = \frac{\hat{\gamma}_{\ell 1}}{\hat{\gamma}_{\ell 1}+\hat{\gamma}_{\ell 2}} \prod_{r=1}^{\ell-1} \frac{\hat{\gamma}_{r2}}{\hat{\gamma}_{r1}+\hat{\gamma}_{r2}}$, for $\ell = 2, ..., N-1$ (the expressions for $\ell = 1$ and $\ell = N$ are special cases).

- However, we do not have a closed form expression for $\mathrm{E}_q\{\mathrm{N}(y_0 \mid Z_\ell, \phi^{-1})\} = \int\int \{(2\pi)^{-1/2}\phi^{1/2} \exp(-0.5\phi(y_0 - Z_\ell)^2)\} \mathrm{N}(Z_\ell \mid \hat{\xi}_{\ell 1}, \hat{\xi}_{\ell 2})\, \mathrm{ga}(\phi \mid \hat{\beta}_1, \hat{\beta}_2)\, \mathrm{d}Z_\ell \mathrm{d}\phi$. Use MC integration, with samples from $\mathrm{ga}(\phi \mid \hat{\beta}_1, \hat{\beta}_2)$, after integrating w.r.t. Z_ℓ to obtain:

$$
\mathrm{E}_q\{\mathrm{N}(y_0 \mid Z_\ell, \phi^{-1})\} = \int \{2\pi(\hat{\xi}_{\ell 2} + \phi^{-1})\}^{-1/2} \mathrm{e}^{-\frac{(y_0 - \hat{\xi}_{\ell 1})^2}{2(\hat{\xi}_{\ell 2}+\phi^{-1})}}\, \mathrm{ga}(\phi \mid \hat{\beta}_1, \hat{\beta}_2)\, \mathrm{d}\phi
$$

Applications of DP mixture models: some references

Dirichlet process mixture models, and their extensions, have largely dominated applied Bayesian nonparametric work, after the technology for their simulation-based model fitting was introduced. Included below is a sample of references categorized by methodological/application area.

- Density estimation, mixture deconvolution, and density regression: West et al. (1994); Escobar and West (1995); Cao and West (1996); Gasparini (1996); Müller et al. (1996); Ishwaran and James (2002); Do, Müller and Tang (2005); Leslie et al. (2007); Lijoi, Mena and Prünster (2007); Taddy and Kottas (2010).

- Generalized linear, and linear mixed, models; methods for longitudinal data analysis: Bush and MacEachern (1996); Kleinman and Ibrahim (1998a,b); Mukhopadhyay and Gelfand (1997); Müller and Rosner (1997); Quintana (1998); Kyung, Gill and Casella (2010); Hannah et al. (2011); Quintana et al. (2016).

Applications of DP mixture models: some references

- Methods for longitudinal cluster analysis and for functional clustering: Ray and Mallick (2006); Bigelow and Dunson (2009); Petrone, Guindani and Gelfand (2009).

- Regression modeling with structured error distributions and/or regression functions: Brunner (1995); Lavine and Mockus (1995); Kottas and Gelfand (2001); Dunson (2005); Kottas and Krnjajić (2009).

- Regression models for survival/reliability data: Kuo and Mallick (1997); Gelfand and Kottas (2003); Merrick et al. (2003); Argiento et al. (2009); De Iorio et al. (2009).

- Models for binary and ordinal data: Basu and Mukhopadhyay (2000); Hoff (2005); Das and Chattopadhyay (2004); Kottas, Müller and Quintana (2005); Shahbaba and Neal (2009); Bao and Hanson (2015); DeYoreo and Kottas (2015, 2018a,b).

Applications of DP mixture models: some references

- Errors-in-variables models; multiple comparisons problems; analysis of selection models: Müller and Roeder (1997); Gopalan and Berry (1998); Lee and Berger (1999).
- ROC data analysis: Erkanli et al. (2006); Hanson, Kottas and Branscum (2008).
- Meta-analysis and nonparametric ANOVA models: Mallick and Walker (1997); Tomlinson and Escobar (1999); Burr et al. (2003); De Iorio et al. (2004); Müller et al. (2004); Müller et al. (2005).
- Mixture models for Markov time series; time series modeling and econometrics applications: Müller, West and MacEachern (1997); Chib and Hamilton (2002); Hirano (2002); Hasegawa and Kozumi (2003); Griffin and Steel (2004); Tang and Ghosal (2007); Di Lucca et al. (2013); Antoniano-Villalobos and Walker (2016); DeYoreo and Kottas (2017); Kalli and Griffin (2018).

Semiparametric random effects models

- Linear random effects models (e.g., Laird and Ware, 1982) are a widely used class of models for repeated measurements,

$$\boldsymbol{y}_i = X_i \boldsymbol{\beta} + Z_i \boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \qquad i = 1, \ldots, n,$$

  where: $\boldsymbol{y}_i$ is the response vector for the i-th subject; $\boldsymbol{\beta}$ is the vector of fixed effects regression parameters; $\boldsymbol{b}_i$ is the vector of random effects; X_i and Z_i are covariate matrices associated with the fixed and random effects, respectively; and $\boldsymbol{\epsilon}_i$ is the vector of observational errors.

- It is common to assume that $\boldsymbol{b}_i$ is independent from $\boldsymbol{\epsilon}_i$, and that $\boldsymbol{\epsilon}_i \sim \mathrm{N}(0, \sigma^2 I)$.

- Furthermore, it is very common to assume that $\boldsymbol{b}_i \sim \mathrm{N}(0, D)$, mostly because of computational convenience.

Semiparametric random effects models

- Consider a special case, the random intercepts model:

$$y_{ij} = \mu + \theta_i + \epsilon_{ij}, \qquad \theta_i \sim \mathrm{N}(0, \tau^2), \qquad \epsilon_{ij} \sim \mathrm{N}(0, \sigma^2),$$

  for $j = 1, \ldots, m_i$ and $i = 1, \ldots, n$.

- A Bayesian formulation of this model also includes priors on μ, τ^2 and σ^2, e.g,

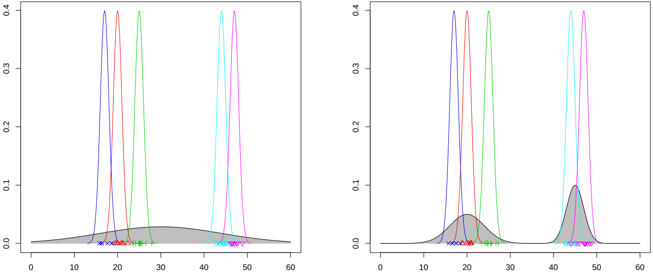
$$\mu \sim \mathrm{N}(\mu_0, \kappa^2) \qquad \sigma^2 \sim \mathrm{IG}(a, b) \qquad \tau^2 \sim \mathrm{IG}(c, d)$$

  (When selecting hyperparameters, recall that an improper prior for σ^2 would be OK, but improper priors for τ^2 are not.)

- When is the assumption of normality for the random effects distribution reasonable?

Random effects distributions

- Normality is, in general, an inappropriate assumption for the random effects distribution.



- Instead, we would often expect the random effects distribution to present multimodalities because of the effects of covariates that have not been included in the model.

Bayesian semiparametric random effects models

- Bayesian semiparametric random effects models have been discussed in Bush and MacEachern (1996), Kleinman and Ibrahim (1998a,b), Mukhopadhyay and Gelfand (1997), Burr and Doss (2005), and Kyung, Gill and Casella (2010), in addition to a number of applied papers.

- General formulation:

$$
\begin{aligned}
\boldsymbol{y}_i \mid \boldsymbol{\beta}, \boldsymbol{b}_i, \sigma^2 &\sim \mathrm{N}(X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i, \sigma^2 I), && i = 1, \dots, n \\
\boldsymbol{b}_i \mid G &\sim G, && i = 1, \dots, n \\
G \mid \alpha, D &\sim \mathrm{DP}(\alpha, \mathrm{N}(0, D)) \\
\boldsymbol{\beta}, \sigma^2, \alpha, D &\sim p(\beta, \sigma^2, \alpha, D)
\end{aligned}
$$

Bayesian semiparametric random intercepts model

- For the random intercepts model:

$$\begin{aligned} y_{ij} \mid \theta_i, \sigma^2 &\sim \mathrm{N}(\theta_i, \sigma^2), & j = 1, \dots, m_i, \quad i = 1, \dots, n \\ \theta_i \mid G &\sim G, & i = 1, \dots, n \\ G \mid \alpha, \mu, \tau^2 &\sim \mathrm{DP}(\alpha, \mathrm{N}(\mu, \tau^2)) \end{aligned}$$

  with hyperpriors for σ^2 and (some of) the DP parameters (α, μ, τ^2) (note that, without loss of generality, we absorbed the intercept μ).

- For $\alpha \to \infty$ we recover the traditional Gaussian random effects model, whereas for $\alpha \to 0$, the model reduces to a parametric model without random effects.
- For values of α in between, the model induces ties among the θ_i.

Density regression using Dirichlet process mixtures

- Two dominant trends in the Bayesian regression literature: seek increasingly flexible regression function models, and accompany these models with general error distributions.

- Typically, Bayesian nonparametric modeling focuses on either the regression function or the error distribution.

- Bayesian nonparametric models for *density regression* (aka *conditional regression*) (West et al., 1994; Müller et al., 1996).
    - Flexible nonparametric mixture modeling for the joint distribution of response(s) and covariates.
    - Inference for the conditional response distribution given covariates.

- Both the response distribution and, implicitly, the regression relationship are modeled nonparametrically, thus providing a flexible framework for the general regression problem.

Density regression 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 $f(y, \mathrm{x})$ of the response y and the vector of covariates x:

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

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

Density regression using Dirichlet process mixtures

- For any grid of values $(y_0, \boldsymbol{x}_0)$, obtain posterior samples for:
  - Joint density $f(y_0, \mathrm{x}_0 \mid G)$, marginal density $f(\mathrm{x}_0 \mid G)$, and therefore, conditional density $f(y_0 \mid \mathrm{x}_0, G)$.
  - Conditional expectation $\mathrm{E}(y \mid \mathrm{x}_0, G)$, which, estimated over grid in x, provides inference for the mean regression relationship.
  - Conditioning in $f(y_0 \mid \mathrm{x}_0, G)$ and/or $\mathrm{E}(y \mid \mathrm{x}_0, G)$ may involve only a portion of vector x.
  - *Inverse inferences*: inference for the conditional distribution of covariates given specified response values, $f(\mathrm{x}_0 \mid y_0, G)$.

- Key features of the modeling approach:
  - Model for both non-linear regression curves **and** non-standard shapes for the conditional response density.
  - Model does not rely on additive regression formulations; it can uncover interactions between covariates that might influence the regression relationship.

Mean regression functional under normal DP mixtures

- Assume a normal DP mixture for the joint response-covariate density (univariate response y, covariate vector $\boldsymbol{x} = (x_1, ..., x_p)$)

$$f(y, \boldsymbol{x} \mid G) = \sum_{\ell=1}^{\infty} \omega_\ell \, \mathrm{N}_{p+1}(y, \boldsymbol{x} \mid \boldsymbol{\mu}_\ell, \Sigma_\ell)$$

- Consider the decomposition of $\boldsymbol{\mu}_\ell = (\mu_\ell^y, \boldsymbol{\mu}_\ell^{\boldsymbol{x}})$ and $\Sigma_\ell = (\Sigma_\ell^y, \Sigma_\ell^{y\boldsymbol{x}}, \Sigma_\ell^{\boldsymbol{x}})$ into components that correspond to the response and covariates.
- Then, $f(y \mid \boldsymbol{x}, G) = \sum_{\ell=1}^{\infty} q_\ell(\boldsymbol{x}) \, \mathrm{N}(y \mid \lambda_\ell(\boldsymbol{x}), \tau_\ell^2)$, where
  - $q_\ell(\boldsymbol{x}) = \omega_\ell \mathrm{N}_p(\boldsymbol{x} \mid \boldsymbol{\mu}_\ell^{\boldsymbol{x}}, \Sigma_\ell^{\boldsymbol{x}}) / \{\sum_{s=1}^{\infty} \omega_s \mathrm{N}_p(\boldsymbol{x} \mid \boldsymbol{\mu}_s^{\boldsymbol{x}}, \Sigma_s^{\boldsymbol{x}})\}$
  - $\lambda_\ell(\boldsymbol{x}) = \mu_\ell^y + \Sigma_\ell^{y\boldsymbol{x}} (\Sigma_\ell^{\boldsymbol{x}})^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_\ell^{\boldsymbol{x}})$ and $\tau_\ell^2 = \Sigma_\ell^y - \Sigma_\ell^{y\boldsymbol{x}} (\Sigma_\ell^{\boldsymbol{x}})^{-1} (\Sigma_\ell^{y\boldsymbol{x}})^T$
- Mean regression function:

$$\mathrm{E}(y \mid \boldsymbol{x}, G) = \sum_{\ell=1}^{\infty} q_\ell(\boldsymbol{x}) \{\beta_{0\ell} + \beta_{1\ell} x_1 + \ldots + \beta_{p\ell} x_p\}$$

where $\beta_{0\ell} = \mu_\ell^y - \Sigma_\ell^{y\boldsymbol{x}} (\Sigma_\ell^{\boldsymbol{x}})^{-1} \boldsymbol{\mu}_\ell^{\boldsymbol{x}}$, and $\beta_{r\ell}$, for $r = 1, ..., p$, are the elements of vector $\Sigma_\ell^{y\boldsymbol{x}} (\Sigma_\ell^{\boldsymbol{x}})^{-1}$.

Synthetic data example

- Simulated data set with a continuous response y, one continuous covariate x_c, and one binary categorical covariate x_d.
    - x_{ci} independent $\mathrm{N}(0,1)$.
    - $x_{di} \mid x_{ci}$ independent $\mathrm{Ber}(\mathrm{probit}(x_{ci}))$.
    - $y_i \mid x_{ci}, x_{di}$ ind. $\mathrm{N}(h(x_{ci}), \sigma_{x_{di}})$, with $\sigma_0 = 0.25$, $\sigma_1 = 0.5$, and

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

- Two sample sizes: $n = 200$ and $n = 2000$.
- DP mixture model with a mixed normal/Bernoulli kernel:

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

with

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

Synthetic data example

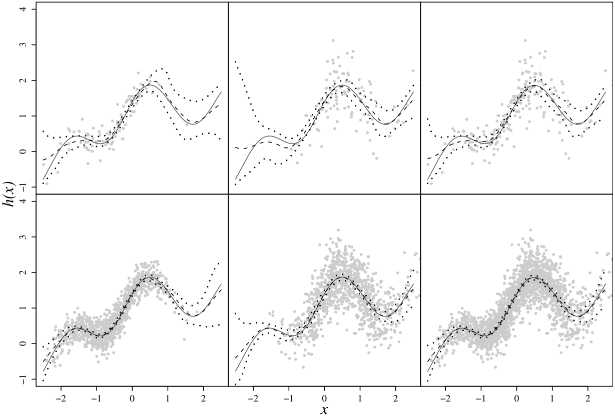


Figure 3.3. Posterior point and 90% interval estimates (dashed and dotted lines) for conditional response expectation $E(y \mid x_c, x_d = 0, G)$ (left panels), $E(y \mid x_c, x_d = 1, G)$ (middle panels), and $E(y \mid x_c, G)$ (right panels). The corresponding data is plotted in grey for the sample of size $n = 200$ (top panels) and $n = 2000$ (bottom panels). The solid line denotes the true curve.

Quantile regression

- In regression settings, the covariates may have effect not only on the location of the response distribution but also on its shape.

- Model-based nonparametric approach to quantile regression.
  - 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:

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

with $G_0(\boldsymbol{\mu}, \Sigma) = \mathrm{N}_{M+1}(\boldsymbol{\mu} \mid \mathrm{m}, V)\mathrm{IW}(\Sigma \mid \nu, S)$.

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

- Key features of the DP mixture modeling framework:
  - Enables simultaneous inference for more than one quantile regression.
  - Allows flexible response distributions **and** non-linear quantile regression relationships.

Quantile regression: 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 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).

Quantile regression: data example

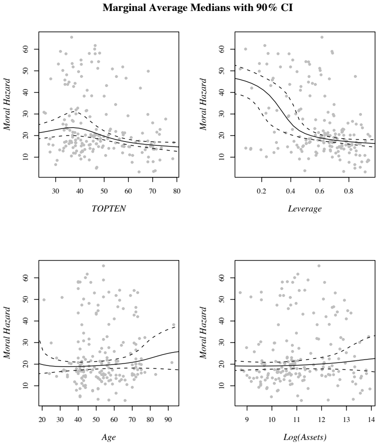


Figure 3.4. Posterior mean and 90% interval estimates for median regression for *MH*5 conditional on each individual covariate. Data scatterplots are shown in grey.

Quantile regression: data example

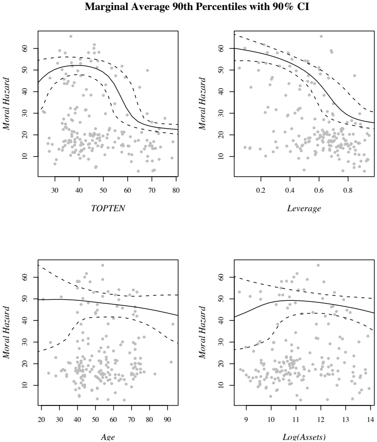


Figure 3.5. Posterior mean and 90% interval estimates for 90th percentile regression for *MH*5 conditional on each individual covariate. Data scatterplots are shown in grey.

Quantile regression: data example

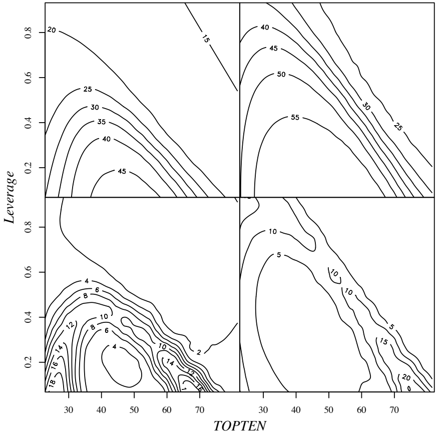


Figure 3.6. Posterior estimates of median surfaces (left column) and 90th percentile surfaces (right column) for *MH*5 conditional on *Leverage* and *TOPTEN*. The posterior mean is shown on the top row and the posterior interquartile range on the bottom.

Quantile regression: data example

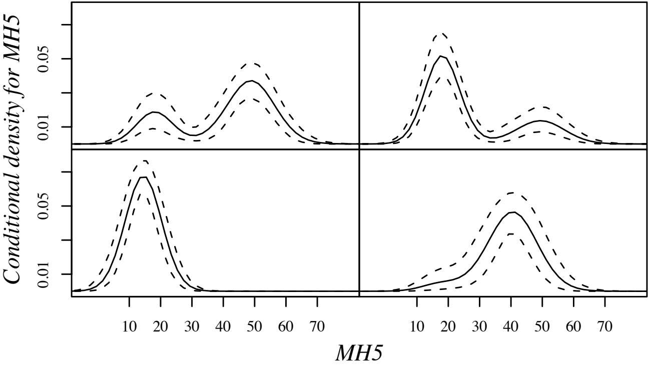


Figure 3.7. Posterior mean and 90% interval estimates for response densities $f(y \mid \boldsymbol{x}_0, G)$ conditional on four combinations of values $\boldsymbol{x}_0$ for the covariate vector (*TOPTEN*, *Leverage*, *Age*, log(*Assets*))

DP mixture density regression: applications

- Regression modeling with categorical responses (Shahbaba & Neal, 2009; Dunson & Bhattacharya, 2011; Hannah et al., 2011; DeYoreo & Kottas, 2015, 2018a,b).
- Functional data analysis through density estimation (Rodriguez et al., 2009).
- Markov switching regression (Taddy & Kottas, 2009), and fully nonparametric quantile regression (Taddy & Kottas, 2010).
- Product partition models with regression on covariates (Müller & Quintana, 2010; Park & Dunson, 2010), and regression modeling with *enriched* DP priors (Wade et al., 2014).
- Nonparametric survival regression (Poynor & Kottas, 2017).
- NPB density autoregression (Heiner & Kottas, 2020).

Modeling for multivariate ordinal data

- Values of k ordinal categorical variables $Y_1, \dots, Y_k$ recorded for n subjects:
  - $C_j \geq 2$: number of categories for the j-th variable, $j = 1, \dots, k$.
  - $n_{\ell_1 \cdots \ell_k}$: number of observations such that

$$\mathrm{Y} = (Y_1, \dots, Y_k) = (\ell_1, \dots, \ell_k).$$

  - $p_{\ell_1 \cdots \ell_k} = \Pr(Y_1 = \ell_1, \dots, Y_k = \ell_k)$ is the classification probability for the $(\ell_1, \dots, \ell_k)$ 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 $\sum_{\ell_1 \cdots \ell_k} n_{\ell_1 \cdots \ell_k} = n$.

Modeling for multivariate ordinal data

- A possible modeling strategy (alternative to log-linear models) involves the introduction of k continuous latent variables $\mathrm{Z} = (Z_1, \dots, Z_k)$ whose joint distribution yields through discretization the classification probabilities for the table cells,

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

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

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

Modeling for multivariate ordinal data

- Richer modeling and inference based on normal DP mixtures for the latent variables Z_i associated with data vectors Y_i, $i = 1, \dots, n$.

- Model $\mathrm{Z}_i \mid G$ i.i.d. f, with $f(\cdot \mid G) = \int N_k(\cdot \mid \boldsymbol{m}, \boldsymbol{S}) \mathrm{d}G(\boldsymbol{m}, \boldsymbol{S})$, where

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

- Advantages of the DP mixture modeling approach:
  - Can accommodate essentially any pattern in k-dimensional contingency tables.
  - Allows local dependence structure to vary accross the contingency table.
  - Implementation does not require random cutoffs (so the complex updating mechanisms for cutoffs are not needed).

Modeling for multivariate ordinal data: data example

- A data set on *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, $\Pr(n^* \geq 4 \mid \text{data}) = 1$.

Modeling for multivariate ordinal data: data example

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)

Modeling for multivariate ordinal data: data example

- Posterior predictive distributions $p(\mathrm{Z}_0 \mid \mathrm{data})$ (see Figure 3.8) – DP mixture version is based on the posterior predictive distribution for corresponding mixing parameter $(\mathrm{m}_0, \mathrm{S}_0)$.
- Inference for the association between the ordinal variables:
  - For example, Figure 3.8 shows posteriors for ρ_0, the correlation coefficient implied in S_0.
  - The probit model does not capture successfully the association of the ordinal variables, since it fails to recognize the clustering suggested by the data (revealed by the DP mixture model).
- Figure 3.9 shows inferences for log-odds ratios,

$$\psi_{ij} = logp_{i,j} + \log p_{i+1,j+1} - \log p_{i,j+1} - \log p_{i+1,j}.$$

- Utility of mixture modeling 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.

Modeling for multivariate ordinal data: data example

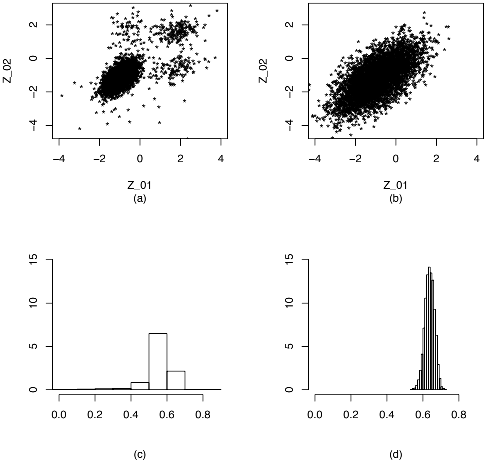


Figure 3.8. For the interrater agreement data, draws from $p(Z_0 \mid \text{data})$ and $p(\rho_0 \mid \text{data})$ under the DP mixture model (panels (a) and (c), respectively) and the probit model (panels (b) and (d), respectively).

Modeling for multivariate ordinal data: data example

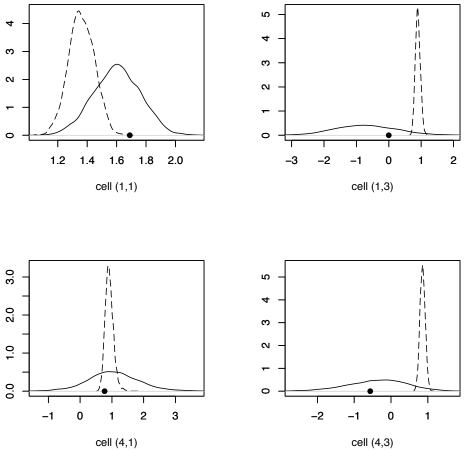


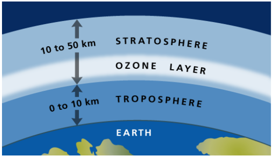
Figure 3.9. For the interrater agreement data, posteriors for four log-odds ratios under the DP mixture model (solid lines) and the probit model (dashed lines). The circles denote the corresponding empirical log-odds ratios.

Nonparametric multivariate ordinal regression

- k ordinal variables $\mathrm{Y} = (Y_1, \dots, Y_k)$, with $y_j \in \{1, \dots, C_j\}$, and p (continuous) covariates $\mathrm{X} = (X_1, \dots, X_p)$.
- Again, $Y_j = \ell$ if-f $\gamma_{j,\ell-1} < Z_j \leq \gamma_{j,\ell}$, for $j = 1, ..., k$, and $\ell = 1, ..., C_j$.
- Now, model the joint distribution of the latent continuous responses, $\mathrm{Z} = (Z_1, \dots, Z_k)$, and the covariates, X, with a multivariate normal DP mixture $\rightarrow$ implies a regression model, $\Pr(\mathrm{Y} \mid \mathrm{x})$, which is a mixture of probit regressions with covariate-dependent weights.
- Large support under fixed cut-offs:
  - for any mixed ordinal-continuous distribution, $p_0(\boldsymbol{x}, \boldsymbol{y})$, that satisfies certain regularity conditions, the prior model assigns positive probability to all Kullback-Leibler (KL) neighborhoods of $p_0(\boldsymbol{x}, \boldsymbol{y})$, as well as to all KL neighborhoods of the implied conditional distribution, $p_0(\boldsymbol{y} \mid \boldsymbol{x})$.

Ozone concentration data example

- Data set comprising 111 measurements of ozone concentration (ppb), wind speed (mph), radiation (langleys), and temperature (degrees Fahrenheit).



- Ozone concentration recorded on continuous scale.
- To construct an ordinal response: define "high" as above 100 ppb, "medium" as $(50, 100]$ ppb, and "low" as less than 50 ppb.
- Comparison of inferences from the model for (Y, X) with those from a DP mixture of normals model for (Z, X).

Ozone concentration data example

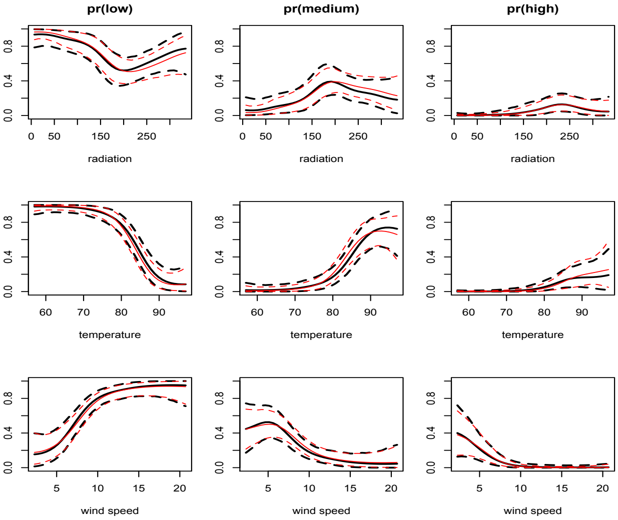


Figure 3.10. Posterior mean (solid) and 95% interval estimates (dashed) for $\Pr(Y = \ell \mid x_m, G)$ (black) compared to $\Pr(\gamma_{\ell-1} < Z \leq \gamma_\ell \mid x_m, G)$ (red).

Ozone concentration data example

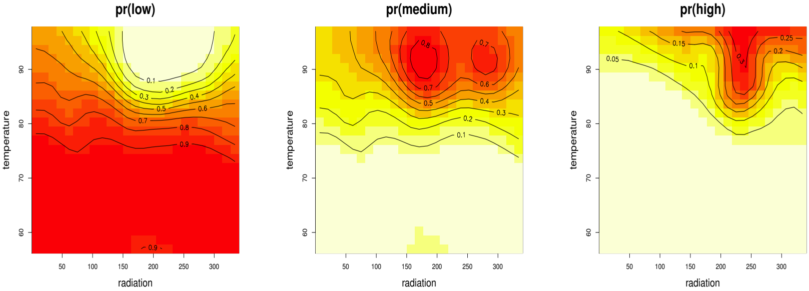


Figure 3.11. Posterior mean estimates for $\Pr(Y = \ell \mid x_1, x_2, G)$, for $\ell = 1, 2, 3$, corresponding to low (left), medium (middle) and high (right). Red represents a value of 1, white represents 0.

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).

- 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).

Definition of Poisson processes on the real line

- 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:
  - For any $t > s \geq 0$, $N(t) - N(s)$ follows a Poisson distribution with mean $\Lambda(t) - \Lambda(s)$.
  - $\mathcal{N}$ 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 $B \subset \mathbb{R}^+$).

- So, from a modeling perspective, the main functional of interest for a NHPP is its intensity function.

Nonparametric inference for Poisson processes

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

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

- **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, a nonparametric prior model for f, with a parametric prior for γ, will induce a semiparametric prior for λ.
- Since γ only scales λ, it is f that determines the shape of the intensity function λ.

Nonparametric inference for Poisson processes

- **Beta DP mixture model** for f:

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

where $\text{Beta}(t \mid \mu, \tau)$ is the Beta density on $(0, T)$ with mean $\mu \in (0, T)$ and scale parameter $\tau > 0$, and $G_0(\mu, \tau) = \text{Uni}(\mu \mid 0, T)$ $\text{IG}(\tau \mid c, \beta)$ with random scale parameter β.

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

- Full Bayesian model:

$$e^{-\gamma} \gamma^n \left\{ \prod_{i=1}^{n} \int \text{Beta}(t_i \mid \mu_i, \tau_i) \text{d}G(\mu_i, \tau_i) \right\} p(\gamma) \text{DP}(G \mid \alpha, G_0(\beta)) p(\alpha) p(\beta)$$

- Reference prior for γ, $p(\gamma) \propto \gamma^{-1}$.

Nonparametric inference for Poisson processes

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

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

  where:

  - $p(\gamma \mid \text{data})$ is a gamma$(n, 1)$ distribution.
  - MCMC is used to sample from $p(\boldsymbol{\theta}, \alpha, \beta \mid \text{data})$.
  - $p(G \mid \boldsymbol{\theta}, \alpha, \beta)$ is a DP with updated parameters (can be sampled as discussed earlier).

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

- Extensions to inference for spatial NHPP intensities, using DP mixtures with bivariate Beta kernels (Kottas and Sansó, 2007).

Data examples

- Example for temporal NHPPs: 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.

- Specification for $\text{DP}(\alpha, G_0(\mu, \tau \mid \beta) = \text{Uni}(\mu \mid 0, T)\text{IG}(\tau \mid 2, \beta))$.
  - $\text{gamma}(a_\alpha, b_\alpha)$ prior for α.
  - 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 possible 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, using two forestry data sets:
  - locations of 62 redwood seedlings in a square of 23 m;
  - locations of 514 maple trees in a 19.6 acre square plot in Lansing Woods, Clinton County, MI.

Data examples

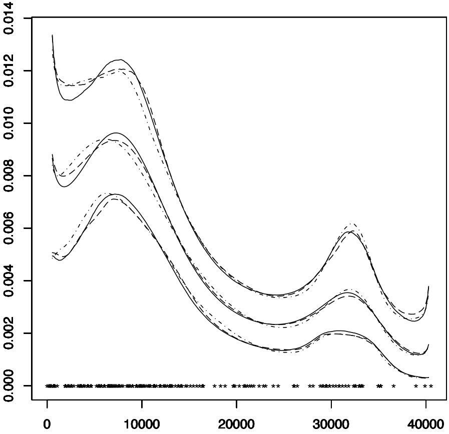


Figure 3.12. Coal-mining disasters data. Posterior point and 95% interval estimates for the intensity function under three prior settings. The observed times of the 191 explosions in mines are plotted on the horizontal axis.

Data examples

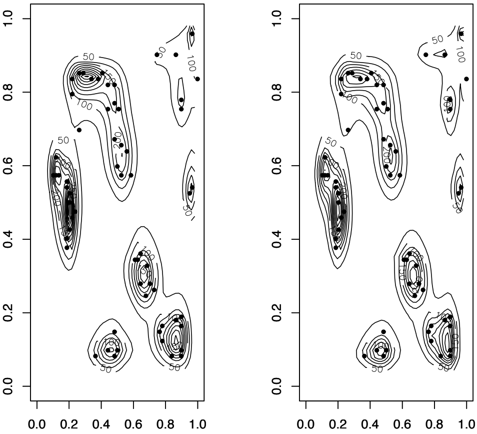


Figure 3.13. Redwood seedlings data. Contour plots of posterior mean intensity estimates under two different priors for α. The dots indicate the locations of the redwood seedlings.

Data examples

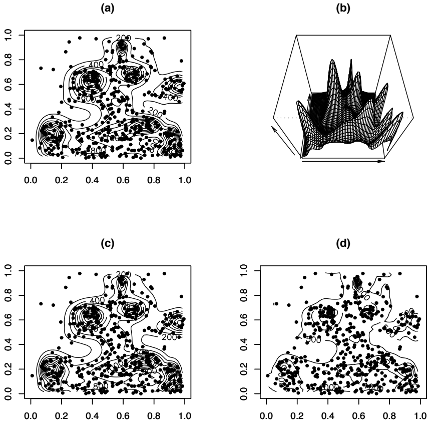


Figure 3.14. Maples data. Panels (a) and (b) include the posterior mean intensity estimate (contour and perspective plot, respectively). Panels (c) and (d) show contour plots for the posterior median and interquartile range intensity estimates, respectively. The dots denote the locations of the maple trees.

Nonparametric modeling for NHPPs: further work

- Applications to neuronal data analysis (Kottas and Behseta, 2010; Kottas et al., 2012).
- Inference for marked Poisson processes (Taddy & Kottas, 2012).
- Dynamic modeling for spatial NHPPs (Taddy, 2010).
- Risk assessment of extremes from spatially dependent environmental time series (Kottas et al., 2012) and from correlated financial markets (Rodriguez et al., 2017).
- Dynamic modeling for time-varying seasonal intensities, with an application to predicting hurricane damage (Xiao et al., 2015).

- More recent work on prior models for the NHPP intensity, based on weighted combinations of Erlang densities (Kim & Kottas, 2020) or Bernstein densities (Zhao & Kottas, 2021).

4. Nonparametric Priors for Dependent Distributions

Nonparametric priors for dependent distributions

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

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

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

- We would like something in between.

Nonparametric priors for dependent distributions

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

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

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

Modeling dependence in collections of random distributions

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

Outline and further references

- DDP priors.

- Hierarchical nonparametric priors for finite collections of distributions: hierarchical DPs (Teh. et al., 2006), which are related to the "analysis of densities" model (Tomlinson and Escobar, 1999); and nested DPs (Rodriguez et al., 2008).

- Spatial DPs (Gelfand et al., 2005; Kottas et al., 2008)

- Two applications of DDP modeling: risk assessment in developmental toxicity studies (Fronczyk and Kottas, 2014a), and inference for dynamic ordinal regression relationships (DeYoreo and Kottas, 2018b).

- However, this is by no means an exhaustive list: order-depedent DDPs (Griffin and Steel, 2006); generalized spatial DP (Duan, Guindani and Gelfand, 2007); kernel stick-breaking processes (Dunson and Park, 2008); dependent Pólya tree regression models (Trippa et al., 2011); stick-breaking autoregressive processes (Griffin and Steel, 2011); dependent normalized completely random measures (Griffin et al., 2013; Lijoi et al., 2014) .....

Definition of the dependent Dirichlet process

- Recall the DP constructive definition: if $G \sim \mathrm{DP}(\alpha, G_0)$, then

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

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

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

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

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

Dependent Dirichlet processes

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

"Common-weights" dependent Dirichlet processes

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

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

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

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

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

“Common-atoms” dependent Dirichlet processes

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

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

  where the θ_{ℓ} are i.i.d. from G_0.

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

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

ANOVA dependent Dirichlet process models

- Consider a space S such that $\mathrm{s} = (s_1, \ldots, s_p)$ corresponds to a vector of categorical variables. For instance, in a clinical setting, G_{s_1,s_2} might correspond to the random effects distribution for patients treated at levels s_1 and s_2 of two different drugs.

- For example, consider the normal mixture

$$y_{s_1,s_2,k} \mid G_{s_1,s_2}, \sigma^2 \sim \int \mathrm{N}(y_{s_1,s_2,k} \mid \eta, \sigma^2) \mathrm{d}G_{s_1,s_2}(\eta), \quad G_{s_1,s_2} = \sum_{h=1}^{\infty} \omega_h \, \delta_{\theta_{h,s_1,s_2}}$$

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

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

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

ANOVA dependent Dirichlet process models

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

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

  where d_{s_1,s_2} is a design vector selecting the appropriate coefficients from η and $G_0 = G_0^m G_0^A G_0^B G_0^{AB}$.

- In practice, just a small number of ANOVA models. If a single component is used, we recover a parametric ANOVA model.
- Rephrasing the ANOVA-DDP model as a DP mixture simplifies posterior simulation.
  - Function `LDDPdensity` in `DPpackage` implements ANOVA-DDP models.

Linear-DDP models

- An analogue of the ANOVA-DDP for continuous covariates. Consider w.l.o.g. a single continuous covariate, x.
- Example: linear-DDP normal mixture model

$$f(y \mid G_x, \sigma^2) = \int \mathrm{N}(y \mid \theta, \sigma^2)\, \mathrm{d}G_x(\theta) \quad \text{with} \quad G_x = \sum_{\ell=1}^{\infty} \omega_\ell\, \delta_{\beta_{0\ell}+\beta_{1\ell}x}$$

  i.e., use common weights and a linear function (rather than a full stochastic process) for the atoms: $\theta_\ell(x) = \beta_{0\ell} + \beta_{1\ell}x$, with $(\beta_{0\ell}, \beta_{1\ell}) \overset{ind.}{\sim} G_0$.

- The model can be written as a DP mixture of normal linear regressions:

$$f(y \mid G_x, \sigma^2) = \sum_{\ell=1}^{\infty} \omega_\ell\, \mathrm{N}(y \mid \beta_{0\ell} + \beta_{1\ell}x, \sigma^2) = \int \mathrm{N}(y \mid \beta_0 + \beta_1 x, \sigma^2)\, \mathrm{d}F(\beta_0, \beta_1)$$

  where $F = \sum_{\ell=1}^{\infty} \omega_\ell\, \delta_{(\beta_{0\ell}, \beta_{1\ell})}$, i.e., $F \sim \mathrm{DP}(\alpha, G_0)$.

- Flexible in terms of non-Gaussian response distributions, but not in terms of regression relationships: $\mathrm{E}(y \mid G_x) = \sum_{\ell=1}^{\infty} \omega_\ell\, (\beta_{0\ell} + \beta_{1\ell}x)$, a mixture of linear regressions, but without local (covariate-dependent) weights.

Fully nonparametric "random effects" models

- As an example, consider modeling the distribution of SAT scores on different schools.
- So, data point y_{ij} corresponds to the SAT score obtained by student $j = 1, \ldots, m_i$ in school $i = 1, \ldots, n$.
- A standard parametric model for the (possibly transformed) SAT scores is the Gaussian random intercepts model:

$$y_{ij} \mid \theta_i, \sigma^2 \overset{ind.}{\sim} \mathrm{N}(y_{ij} \mid \theta_i, \sigma^2), \qquad \theta_i \mid \mu, \tau^2 \overset{i.i.d.}{\sim} \mathrm{N}(\mu, \tau^2)$$

  with hyperpriors assigned to μ and τ^2. Here, θ_i is the school-specific random effect.

- But, what if the distributions of scores appear to be non-Gaussian?
- Can we develop prior models that allow flexible SAT score distributions **and** general dependence structures to borrow strength across the SAT score distributions over all schools?

Fully nonparametric "random effects" models

- DP mixture models with school-specific mixing distributions:

$$y_{ij} \mid G_i, \sigma^2 \overset{ind.}{\sim} \int \mathrm{N}(y_{ij} \mid \theta, \sigma^2) \, \mathrm{d}G_i(\theta)$$

$$G_i \mid \alpha_i, G_0 \overset{ind.}{\sim} \mathrm{DP}(\alpha_i, G_0)$$

- Dependence? borrowing of strength?
  - introduce dependence through a hierarchical prior for the α_i and/or parameters ψ_i of G_0? restrictive, weak form of dependence
  - a parametric G_0 is in general restrictive

- Use a nonparametric prior structure for G_0
  - analysis of densities model: $G_0 \sim$ DP mixture prior
  - hierarchical DP (HDP): $G_0 \sim$ DP prior
  - nested DP (NDP): $G_0 =$ DP

(typically, $\alpha_i \equiv \alpha$ for the HDP; we must have $\alpha_i \equiv \alpha$ for the NDP).

Analysis of densities model

- School-specific mixing distributions arising conditionally independent from a DP with a DP mixture prior for its centering distribution:

$$
\begin{aligned}
y_{ij} \mid G_i, \sigma^2 &\overset{ind.}{\sim} \int \mathrm{N}(y_{ij} \mid \theta, \sigma^2) \, \mathrm{d}G_i(\theta) \\
G_i \mid \alpha_i, G_0 &\overset{ind.}{\sim} \mathrm{DP}(\alpha_i, G_0) \\
G_0(u \mid F, \tau^2) = \int \mathrm{N}(u \mid \mu, \tau^2) \, \mathrm{d}F(\mu), &\quad F \sim \mathrm{DP}(\alpha_0, F_0)
\end{aligned}
$$

- The DP mixture for G_0 encourages similar (though not identical) G_i, and thus grouping of the SAT score distributions.
- Note that the mixing distributions G_i have different atoms and different weights (even under $\alpha_i \equiv \alpha$).
- The model structure is reminiscent of the Gaussian random effects model, but it is built at the level of the distributions.

Hierarchical Dirichlet processes

- Consider again the example with SAT scores from different schools.
- Hierarchical Dirichlet process (HDP) mixture models estimate the school-specific distribution by identifying latent classes of students that appear (possibly with different frequencies) in all schools.
- Let

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

- Conditionally on G_0, the mixing distribution for each school is an independent realization from the $\mathrm{DP}(\alpha, G_0)$
  - dependence across schools is introduced, since they all share the same baseline distribution G_0.
- As with the analysis of densities model, we recognize a random effects model structure for distributions.

Hierarchical Dirichlet processes

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

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

- Note that the HDP resembles the structure of a common-atoms DDP prior model.

- The weights assigned to the atoms are *not* independent. Intuitively, if ϕ_ℓ has a large weight ω_ℓ under G_0, then the weight $\pi_{\ell i}$ under G_i will likely be large for every i.
- Indeed, $\boldsymbol{\pi}_i \mid \boldsymbol{\omega} \sim \mathrm{DP}(\alpha, \boldsymbol{\omega})$, where $\boldsymbol{\pi}_i = (\pi_{1i}, \pi_{2i}, \ldots)$ and $\boldsymbol{\omega} = \{\omega_\ell : \ell = 1, 2, ...\}$ (see the next page), such that $\mathrm{E}(\pi_{\ell i} \mid \boldsymbol{\omega}) = \omega_\ell$.

Hierarchical Dirichlet processes

- Assume H is a continuous distribution on $\mathbb{R}$.

- Consider a partition $(A_1, ..., A_r)$ of $\mathbb{R}$, and let $K_s = \{\ell : \phi_\ell \in A_s\}$, for $s = 1, ..., r$, such that $(K_1, ..., K_r)$ is a partition of $\mathbb{Z}^+ = \{1, 2, ...\}$. (Since H is continuous, the ϕ_ℓ are distinct a.s., and therefore there is an one-to-one correspondence between the partitions of $\mathbb{R}$ and $\mathbb{Z}^+$.)

- Now, $(G_i(A_1), ..., G_i(A_r)) \mid G_0 \sim \text{Dirichlet}(\alpha G_0(A_1), ..., \alpha G_0(A_r))$, for each i, that is,

$$\left(\sum_{\ell \in K_1} \pi_{\ell i}, \ldots, \sum_{\ell \in K_r} \pi_{\ell i} \right) \mid \boldsymbol{\omega} \sim \text{Dirichlet} \left(\alpha \sum_{\ell \in K_1} \omega_\ell, \ldots, \alpha \sum_{\ell \in K_r} \omega_\ell \right)$$

  for any partition $(K_1, ..., K_r)$ of $\mathbb{Z}^+$. Hence, $\boldsymbol{\pi}_i \mid \boldsymbol{\omega} \sim \text{DP}(\alpha, \boldsymbol{\omega})$, where the centering DP distribution $\boldsymbol{\omega}$ is a distribution on $\mathbb{Z}^+$.

Hierarchical Dirichlet processes

- Using the previous result for partition $(K_1 = \{1, ..., \ell - 1\}, K_2 = \{\ell\}, K_3 = \{\ell + 1, \ell + 2, ...\})$, we have:
  - $(\sum_{s=1}^{\ell-1} \pi_{si}, \pi_{\ell i}, \sum_{s=\ell+1}^{\infty} \pi_{si}) \mid \boldsymbol{\omega} \sim \text{Dirichlet}(\alpha \sum_{s=1}^{\ell-1} \omega_s, \alpha\omega_\ell, \alpha \sum_{s=\ell+1}^{\infty} \omega_s)$
  - and, using Dirichlet distribution properties, $\pi^*_{\ell i} = (1 - \sum_{s=1}^{\ell-1} \pi_{si})^{-1} \pi_{\ell i}$ follows, conditional on $\boldsymbol{\omega}$, a Beta$(\alpha\omega_\ell, \alpha(1 - \sum_{s=1}^{\ell} \omega_s))$ distribution.

- Therefore, for each i, the $\pi_{\ell i}$ admit a stick-breaking representation: $\pi_{1i} = \pi^*_{1i}$ and $\pi_{\ell i} = \pi^*_{\ell i} \prod_{s=1}^{\ell-1}(1 - \pi^*_{si})$, for $\ell \geq 2$, based on the Beta distributed variables $\pi^*_{\ell i}$.
  - This structure can be used to obtain $\text{E}(\pi_{\ell i} \mid \boldsymbol{\omega}) = \omega_\ell$.

- An MCMC sampler can be devised for posterior simulation by composing two Pólya urns, one built from (α, G_0) and one from (β, H). The resulting MCMC algorithm is similar to the marginal sampler for DP mixture models, but bookkeeping is harder.

Nested Dirichlet Processes

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

$$G_i \mid Q \overset{ind.}{\sim} Q = \sum_{k=1}^{\infty} \omega_k \delta_{G_k^*} \qquad G_k^* = \sum_{\ell=1}^{\infty} \pi_{\ell k} \delta_{\theta_{\ell k}},$$

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

Nested Dirichlet Processes

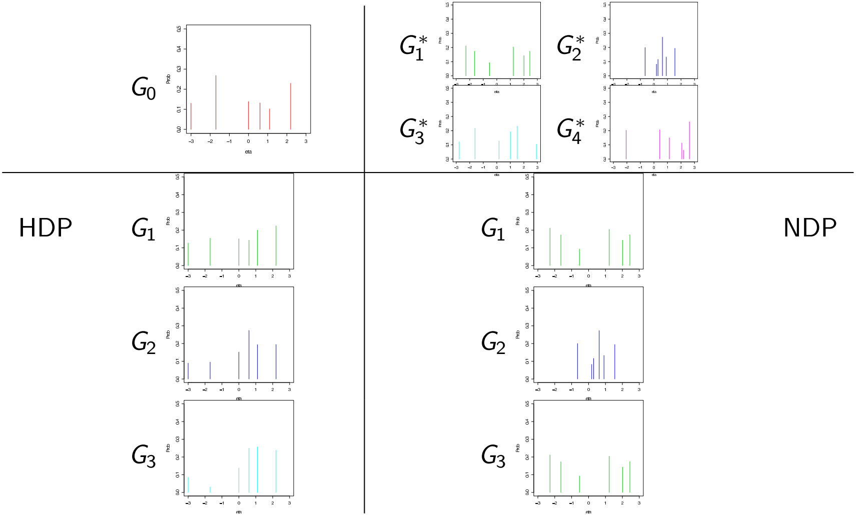
- So, in this case, $G_i \mid Q \overset{ind.}{\sim} Q$, with $Q \sim \mathrm{DP}(\alpha, \mathrm{DP}(\beta, H))$.

- Note that the NDP generates two layers of clustering: states, and hospitals within groups of states. However, groups of states are conditionally independent from each other.

- The HDP vs. the NDP
  - Under the HDP, $\Pr(G_i = G_{i'}) = 0$; same atoms, but different weights for G_i and $G_{i'}$; clustering only for observations.
  - Under the NDP, we have either $G_i = G_{i'}$ or entirely different G_i and $G_{i'}$; if $G_i = G_{i'}$, observations from groups i and i' can be clustered together; clustering on both observations and distributions.

- Fixing an issue with the NDP → latent nested processes (Camerlenghi et al., 2019)

The HDP vs. the NDP



Linear combinations of realizations from independent DPs

- Structured hierarchical model through linear combinations of realizations from independent DPs:

$$G_i = \varepsilon_i \, H_0 + (1 - \varepsilon_i) \, H_i \qquad i = 1, ..., n$$

$$H_i \overset{ind.}{\sim} \text{DP}(\alpha_i, F_0) \qquad i = 0, 1, ..., n$$

- Special case of the model with $\varepsilon_i \equiv \varepsilon$.
- The prior for the ε_i (or for ε) includes point masses at 0 and at 1.

- H_0 is a common component for all distributions G_i, whereas the H_i are idiosyncratic components.

- Under this model, $G_i = G_{i'}$ if-f $\varepsilon_i = \varepsilon_{i'} = 1$, in which case we have $G_i = G_{i'} = H_0$.

Spatial Dirichlet process models

- Spatial data modeling: based on **Gaussian processes** (distributional assumption) and **stationarity** (assumption on the dependence structure).

- Basic model for a spatial random field $\mathrm{Y}_D = \{Y(s) : s \in D\}$, with $D \subseteq R^d$:

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

  - $\mu(s)$ is a mean process, e.g., $\mu(s) = x'(s)\beta$.
  - $\theta(s)$ is a spatial process, typically, a mean 0 isotropic Gaussian process, i.e., $\mathrm{Cov}(\theta(s_i), \theta(s_j) \mid \sigma^2, \phi) = \sigma^2 \rho_\phi(||s_i - s_j||) = \sigma^2 (H(\phi))_{i,j}$
  - $\epsilon(s)$ is a pure error (nugget) process, e.g., $\epsilon(s)$ i.i.d. $N(0, \tau^2)$.

- Induced model for observed sample (**point referenced spatial data**), $\mathrm{Y} = (Y(s_1), \ldots, Y(s_n))$, at sites $\mathrm{s}^{(n)} = (s_1, \ldots, s_n)$ in D

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

Spatial Dirichlet process models

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

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

- Temporal replications available in various applications, e.g., in epidemiology, environmental contamination, and weather modeling.

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

Spatial Dirichlet process models

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

- Hence, the spatial DP is defined as a random process over D

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

  which is centered at G_{0D}.

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

Spatial Dirichlet process models

- Key property: if

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

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

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

Spatial Dirichlet process models

- For stationary G_{0D}, the smoothness of realizations from SDP(α, G_{0D}) is determined by the choice of the covariance function of G_{0D}.
  - For instance, if G_{0D} produces a.s. continuous realizations, then $G^{(s)} - G^{(s')} \to 0$ a.s. as $||s - s'|| \to 0$.
  - We can learn about $G^{(s)}$ more from data at neighboring locations than from data at locations further away (as in usual spatial prediction).

- Random process G_D is centered at a stationary Gaussian process, but it is **nonstationary**, it has **nonconstant variance**, and it yields **non-Gaussian** finite dimensional distributions.

- More general spatial DP models?
  - Allow weights to change with spatial location, i.e., allow realization at location s to come from a different surface than that for the realization at location s' (Duan, Guindani and Gelfand, 2007).

Spatial Dirichlet process models

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

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

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

$$f\left(\text{Y} \mid G^{(n)}, \beta, \tau^2\right) = \int \text{N}_n\left(\text{Y} \mid X'\beta + \theta, \tau^2 I_n\right) \text{d}G^{(n)}\left(\theta\right),$$

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

Spatial Dirichlet process models

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

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

with hyperpriors for β, τ^2, α,σ^2, and ϕ.

- Posterior inference using standard MCMC techniques for DP mixtures — extensions to accommodate missing data — methods for prediction at new spatial locations.

Data example

- Precipitation data from the Languedoc-Rousillon region in southern France.
- Data were discussed, for example, in Damian, Sampson and Guttorp (2001).
  - Original version of the dataset includes 108 altitude-adjusted 10-day aggregated precipitation records for the 39 sites in Figure 4.1.
- 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 4.1, specifically, sites s_4, s_{35}, s_{29}, s_{30}, s_{13}, s_{37}, and refitted the model using only the data from the remaining 33 sites.

Data example

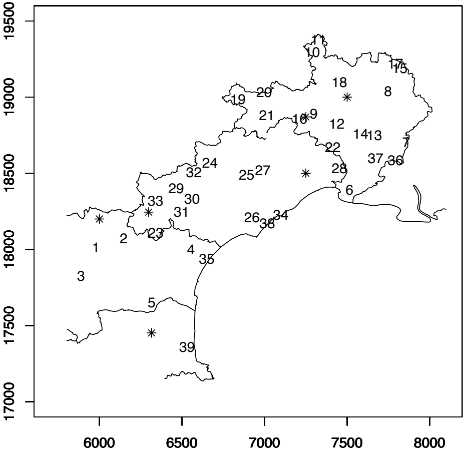


Figure 4.1: Geographic map of the Languedoc-Roussillon region in southern France.

Data example

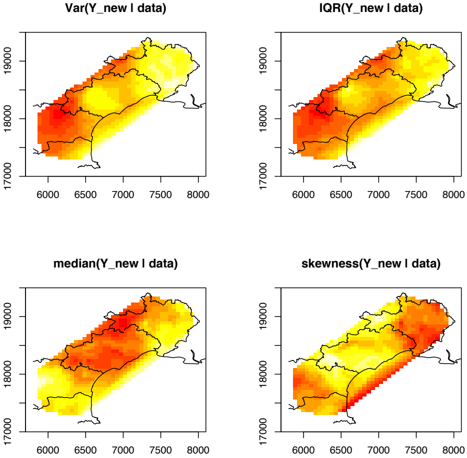


Figure 4.2: French precipitation data. Image plots based on functionals of posterior predictive distributions at observed sites and a number of new sites (darker colors correspond to smaller values).

Data example

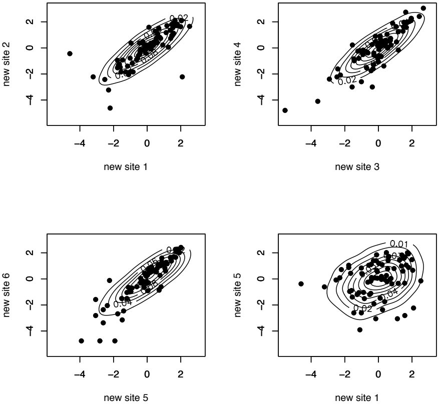


Figure 4.3: French precipitation data. Bivariate posterior predictive densities for pairs of sites (s_4, s_{35}), (s_{29}, s_{30}), (s_{13}, s_{37}) and (s_4, s_{13}) based on model fitted to data after removing sites s_4, s_{35}, s_{29}, s_{30}, s_{13} and s_{37} (overlaid on data observed at the corresponding pairs of sites in the full dataset).

DDP modeling for developmental toxicity studies

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

Developmental toxicology data

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

Developmental toxicology data

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

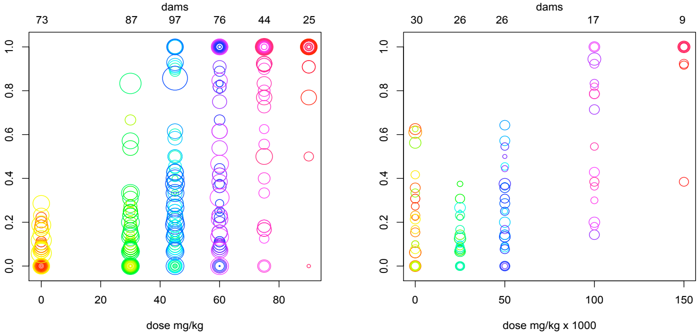


Figure 4.4: 2,4,5-T data (left) and DEHP data (right). Each circle is for a particular dam, the size of the circle is proportional to the number of implants, and the coordinates of the circle are the toxin level and the proportion of combined negative outcomes.

Objectives of DDP modeling

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

DDP mixture model formulation

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

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

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

DDP mixture model formulation

- DDP mixture of Binomial distributions:

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

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

- MCMC posterior simulation using blocked Gibbs sampling.

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

DDP mixture model formulation

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

- Linear mean function for the DDP centering GP enables connections with parametric models, and is key for flexible inference about the dose-response relationship.

Dose-response curve

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

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

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

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

Data examples: 2,4,5-T data

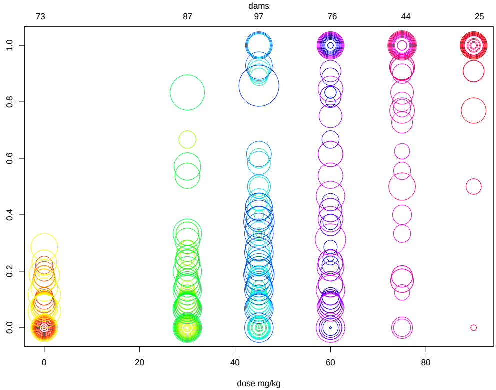


Figure 4.5: 2,4,5-T data. Data set from a developmental toxicity study regarding the effects of the herbicide 2,4,5-trichlorophenoxiacetic (2,4,5-T) acid.

Data examples: 2,4,5-T data

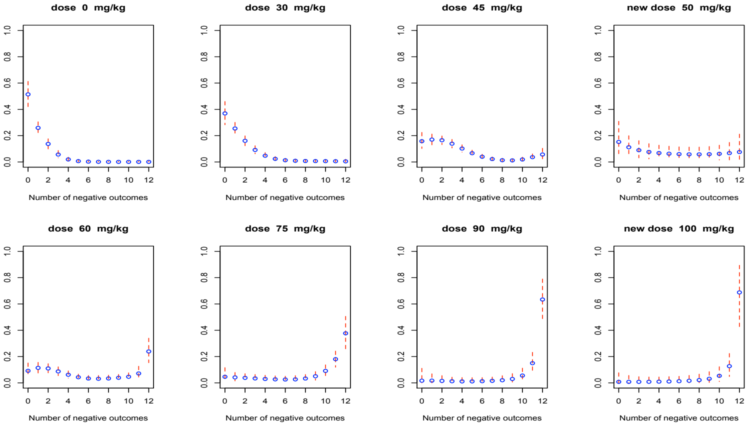


Figure 4.6: 2,4,5-T data. For the 6 observed and 2 new doses, posterior mean estimates (denoted by "o") and 90% uncertainty bands (red) for $f(z \mid m = 12, G_x)$.

Data examples: 2,4,5-T data

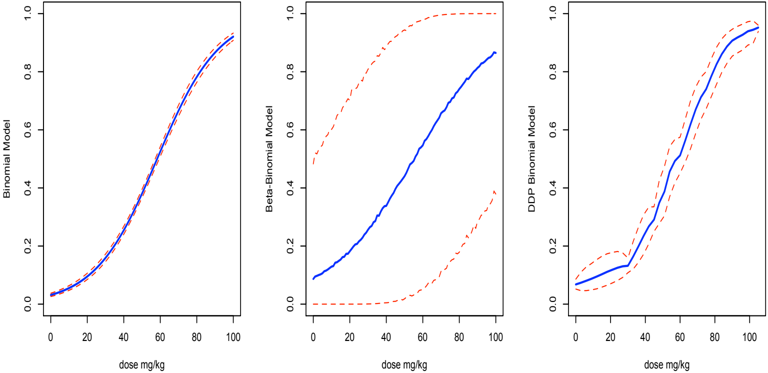


Figure 4.7: 2,4,5-T data. Posterior mean estimate and 90% uncertainty bands for the dose-response curve under a Binomial-logistic model (left), a Beta-Binomial model (middle), and the DDP Binomial mixture model (right).

Data examples: DEHP data

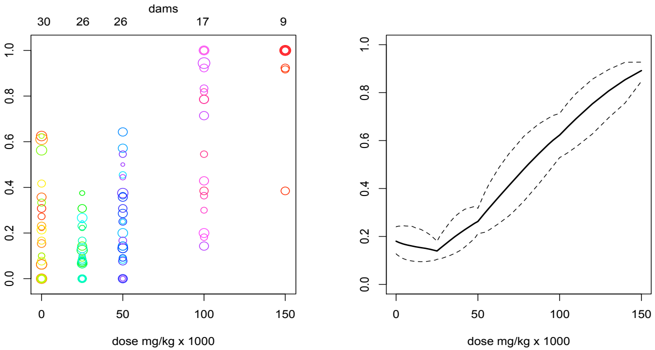


Figure 4.8: DEHP data. Left panel: data from an experiment that explored the effects of diethylhexalphthalate (DEHP), a commonly used plasticizing agent. Right panel: Posterior mean estimate and 90% uncertainty bands for the dose-response curve; the dip at small toxin levels may indicate a hormetic dose-response relationship.

Modeling for multicategory classification responses

Full version of the DEHP data

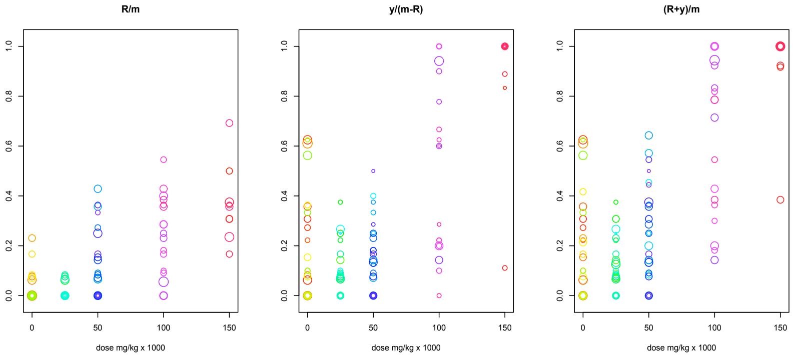


Figure 4.9: Clustered categorical responses: for the j-th dam at dose x_i, R_{ij} resorptions and prenatal deaths, $R_{ij} \leq m_{ij}$ (left panel), and y_{ij} malformations among the live pups, $y_{ij} \leq m_{ij} - R_{ij}$ (middle panel). The right panel plots the combined negative outcomes, $R_{ij} + y_{ij} \leq m_{ij}$, as in Figure 4.8.

Modeling for multicategory classification responses

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

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

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

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

Dose-response curves

- Probability of embryolethality:

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

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

- Probability of malformation:

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

- Combined risk function:

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

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

DEHP data (full version)

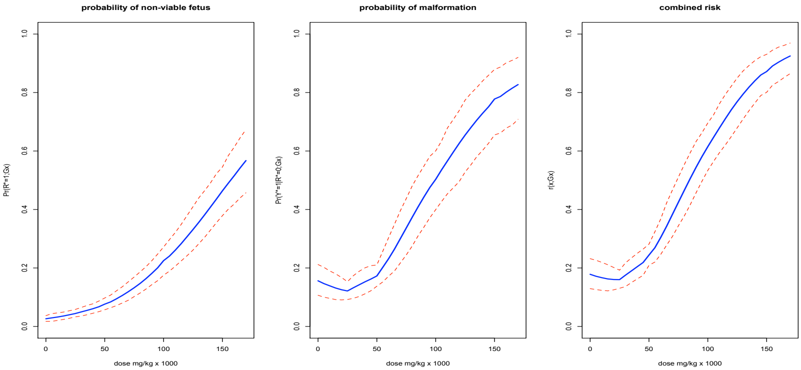


Figure 4.10: DEHP data. Posterior mean estimates and 90% uncertainty bands for the three dose-response curves. The model identifies the malformation endpoint as the sole contributor to the hormetic shape of the combined risk function.

Density regression for ordinal responses

- Recall the nonparametric ordinal regression model from Notes 2.
- Density regression approach: focus on applications, including problems in ecology and the environmental sciences, where it is natural/necessary to model the joint stochastic mechanism for the response(s) and covariates.
- k ordinal variables $\boldsymbol{Y} = (Y_1, \dots, Y_k)$, with $y_j \in \{1, \dots, C_j\}$, and p (continuous) covariates $\boldsymbol{X} = (X_1, \dots, X_p)$.
- Assume $Y_j = \ell$ if-f $\gamma_{j,\ell-1} < Z_j \leq \gamma_{j,\ell}$, for $j = 1, ..., k$, and $\ell = 1, ..., C_j$ (with $\gamma_{j,0} = -\infty$ and $\gamma_{j,C_j} = \infty$).
- Now, model the joint distribution of the latent continuous responses, $\boldsymbol{Z} = (Z_1, \dots, Z_k)$, and the covariates, $\boldsymbol{X}$, with a multivariate normal DP mixture $\rightarrow$ implies a regression model, $\Pr(\boldsymbol{Y} \mid \boldsymbol{x})$, which is a mixture of probit regressions with covariate-dependent weights.

Density regression for ordinal responses

- DP mixture model for $f(\boldsymbol{z}, \boldsymbol{x})$:

$$f(\boldsymbol{z}, \boldsymbol{x} \mid G) = \int \mathrm{N}(\boldsymbol{z}, \boldsymbol{x} \mid \boldsymbol{\mu}, \Sigma) \, \mathrm{d}G(\boldsymbol{\mu}, \Sigma), \quad G \mid \alpha, \boldsymbol{\psi} \sim \mathrm{DP}(\alpha, G_0(\cdot \mid \boldsymbol{\psi}))$$

- Implied regression functions provide a nonparametric extension of probit regression (with random covariates):

$$\Pr(\boldsymbol{Y} = (l_1, \dots, l_k) \mid \boldsymbol{x}, G) = \sum_{r=1}^{\infty} w_r(\boldsymbol{x}) \int_{\gamma_{k,l_k-1}}^{\gamma_{k,l_k}} \cdots \int_{\gamma_{1,l_1-1}}^{\gamma_{1,l_1}} \mathrm{N}(\boldsymbol{z} \mid m_r(\boldsymbol{x}), S_r) \mathrm{d}\boldsymbol{z}$$

  - with covariate dependent weights $w_r(\boldsymbol{x}) \propto p_r \, \mathrm{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_r^x, \Sigma_r^{xx})$
  - and covariate dependent probabilities, where $m_r(\boldsymbol{x}) = \boldsymbol{\mu}_r^z + \Sigma_r^{zx}(\Sigma_r^{xx})^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_r^x)$ and $S_r = \Sigma_r^{zz} - \Sigma_r^{zx}(\Sigma_r^{xx})^{-1}\Sigma_r^{xz}$

Density regression for ordinal responses

- The normal mixture kernel can accommodate continuous covariates, as well as ordinal categorical covariates.

- The prior model has large support under fixed cutoffs.
    - For any mixed ordinal-continuous distribution, $p_0(\boldsymbol{x}, \boldsymbol{y})$, that satisfies certain regularity conditions, the prior assigns positive probability to all Kullback-Leibler (KL) neighborhoods of $p_0(\boldsymbol{x}, \boldsymbol{y})$, as well as to all KL neighborhoods of the implied conditional distribution, $p_0(\boldsymbol{y} \mid \boldsymbol{x})$.

- More flexible ordinal regression relationships **and** simpler posterior simulation (due to fixed cutoffs) than parametric models.

- Posterior simulation: given the continuous latent responses, we can use MCMC methods for normal DP mixture models (the only extra step involves imputing the latent variables).

Extension to dynamic ordinal regression modeling

- Focusing on a univariate ordinal response, we seek to extend to a model for $\mathrm{Pr}_t(Y \mid \boldsymbol{x})$, for $t \in \mathcal{T} = \{1, 2, \dots\}$
- Build on the earlier framework by extending to a prior model for $\{f(\boldsymbol{z}, \boldsymbol{x} \mid G_t) : t \in \mathcal{T}\}$, and thus for $\{\mathrm{Pr}(Y \mid \boldsymbol{x}, G_t) : t \in \mathcal{T}\}$
- Motivating application: data from NMFS on female Chilipepper rockfish collected between 1993 and 2007 along the coast of California
  - sample sizes per year range from 37 to 396, with no data available for three years (2003, 2005 and 2006)
  - three ordinal levels for maturity: immature (1), pre-spawning mature (2), and post-spawning mature (3)
  - length measured in millimeters
  - age recorded on an ordinal scale: age j implies the fish was between j and $j+1$ years of age (data range: 1 to 25) → incorporate age into the model in the same fashion with the maturity variable.

Rockfish data

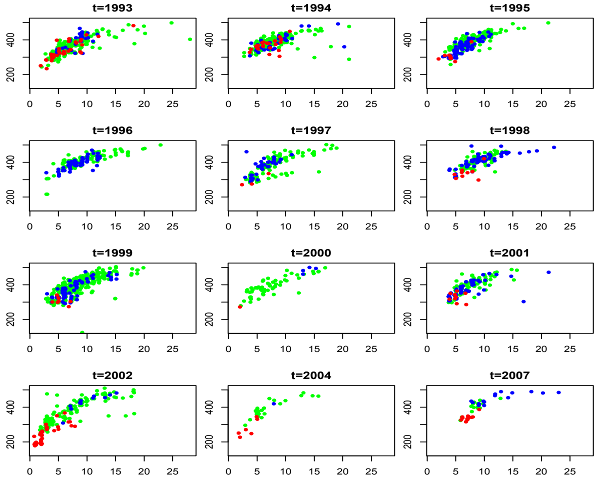


Figure 4.11: Bivariate plots of length versus age at each year of data, with data points colored according to maturity level: red level 1; green level 2; blue level 3.

DDP model extension

- To retain model properties at each t, use DDP prior for $\{G_t : t \in \mathcal{T}\}$
- Time-dependent weights and atoms:

$$f(\boldsymbol{z},\boldsymbol{x} \mid G_t) = \sum_{r=1}^{\infty} \left\{ (1-\beta_{r,t}) \prod\nolimits_{m=1}^{r-1} \beta_{m,t} \right\} \mathrm{N}(\boldsymbol{z},\boldsymbol{x} \mid \boldsymbol{\mu}_{r,t}, \Sigma_r)$$

- Vector autoregressive model for the $\{\boldsymbol{\mu}_{r,t} : t \in \mathcal{T}\}$
  - $\boldsymbol{\mu}_{r,t} \mid \boldsymbol{\mu}_{r,t-1}, \Theta, \boldsymbol{m}, \boldsymbol{V} \sim \mathrm{N}(\boldsymbol{m} + \Theta\boldsymbol{\mu}_{r,t-1}, \boldsymbol{V})$
  - $\Sigma_r \mid \nu, \boldsymbol{D} \overset{i.i.d.}{\sim} \mathrm{IW}(\nu, \boldsymbol{D})$
  - hyperpriors for $(\Theta, \boldsymbol{m}, V)$ and for $\boldsymbol{D}$

DDP model extension

- Stochastic process with beta$(\alpha, 1)$ marginals:

$$\mathcal{B} = \left\{ \beta_t = \exp\left(-\frac{\zeta^2 + \eta_t^2}{2\alpha} \right) : t \in \mathcal{T} \right\}$$

  where $\zeta \sim \mathrm{N}(0,1)$ and, independently, $\{\eta_t : t \in \mathcal{T}\}$ arises from a time series model with $\mathrm{N}(0,1)$ marginals

- Build model for the $\{\beta_{r,t} : t \in \mathcal{T}\}$ from $\beta_{r,t} = \exp\{-(\zeta_r^2 + \eta_{r,t}^2)/(2\alpha)\}$
  - $\zeta_r \overset{ind.}{\sim} \mathrm{N}(0,1)$
  - AR(1) process for $\{\eta_{r,t} : t \in \mathcal{T}\}$: $\eta_{r,t} \mid \eta_{r,t-1}, \phi \sim \mathrm{N}(\phi\eta_{r,t-1}, 1 - \phi^2)$ with $|\phi| < 1$ (and $\eta_{r,1} \overset{ind.}{\sim} \mathrm{N}(0,1)$)

- Different types of correlations can be studied: correlation of the time-dependent stick-breaking weights, and corr$(G_t(A), G_{t+1}(A))$, for any subset A in the support of the G_t .

Rockfish data

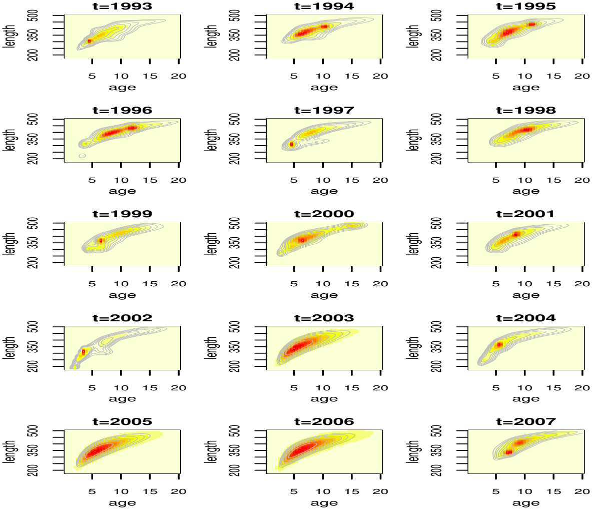


Figure 4.12: Posterior mean estimates for $f(\text{age}, \text{length})$.

Rockfish data

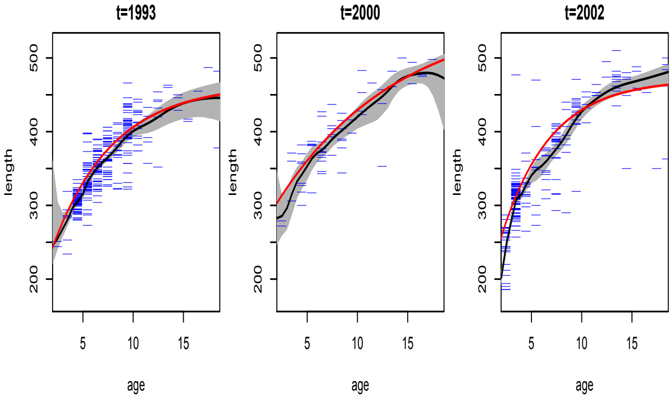


Figure 4.13: Posterior mean and 95% interval bands for the expected value of length over (continuous) age, across three years. Overlaid are the data (in blue) and the estimated von Bertalanffy growth curves (in red).

Rockfish data

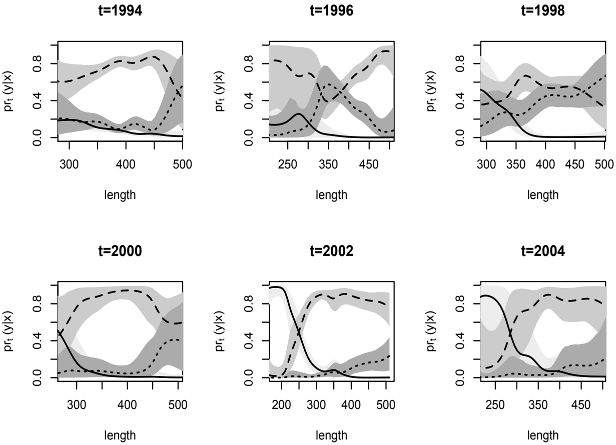


Figure 4.14: Posterior mean and 95% interval bands for the maturation probability curves associated with length: immature (solid); pre-spawning mature (dashed); post-spawning mature (dotted).

Rockfish data

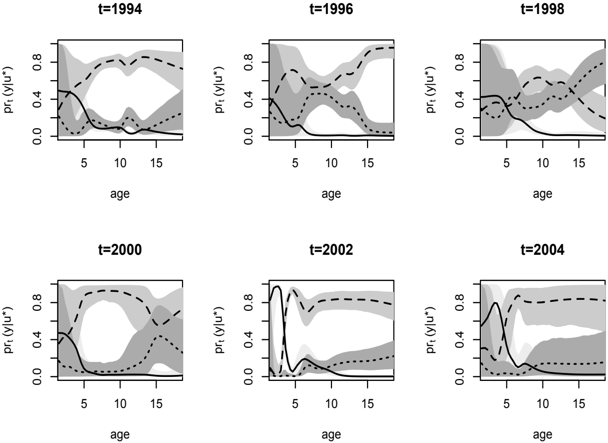


Figure 4.15: Posterior mean and 95% interval bands for the maturation probability curves associated with age: immature (solid); pre-spawning mature (dashed); post-spawning mature (dotted).

Rockfish data

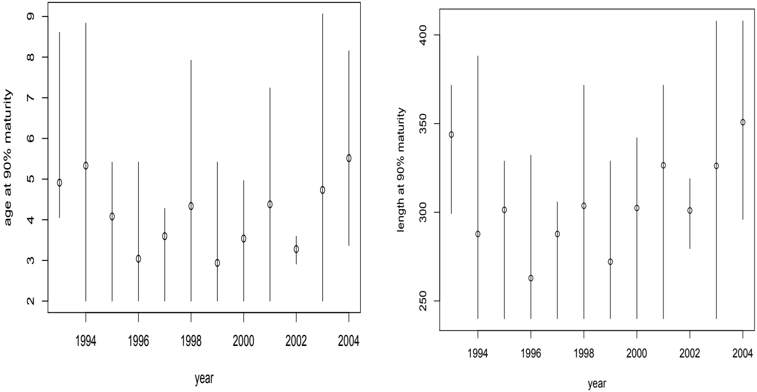


Figure 4.16: Posterior mean and 90% intervals for the smallest value of age above 2 years at which probability of maturity first exceeds 0.9 (left), and similar inference for length (right).

Rockfish data

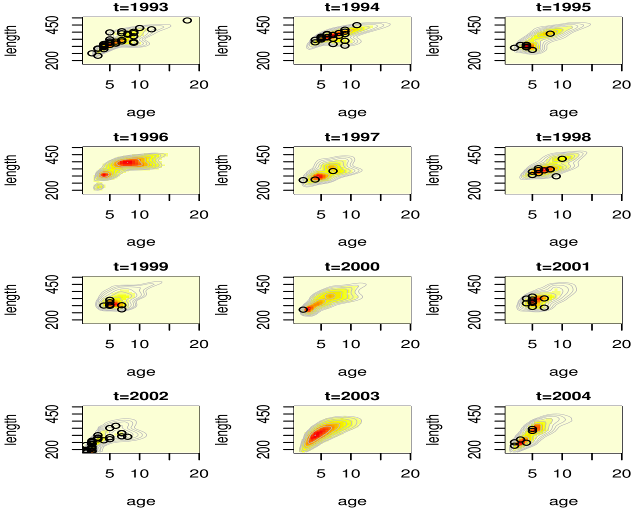


Figure 4.17: Posterior mean estimates for $f(\text{age}, \text{length} \mid Y = 1)$, with corresponding data overlaid.

Rockfish data

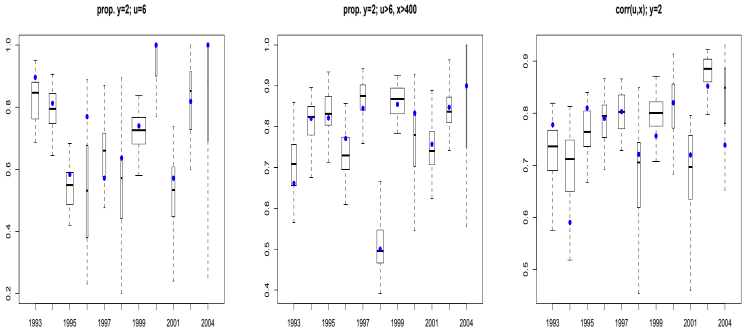


Figure 4.18: Results from posterior predictive model checking. Proportion of age $= 6$ pre-spawning mature fish (left), proportion of age ≥ 7, and length > 400 mm pre-spawning mature fish (middle), and sample correlation between length and age for pre-spawning mature fish. The blue circles in the left and middle panels denote the actual data proportions, and in the right panel the data-based correlation.

Conclusions

- Bayesian nonparametric methods free the data analyst from customary parametric modeling restrictions yielding more general inference and more reliable predictions.

- A broad research field that involves: theoretical work on probability models for spaces of distributions and functions; methodological work on incorporating nonparametric priors into statistical models; computational work on posterior simulation algorithms; and applications.

- About 50 years of history by now, but still going strong!

Many thanks!

- PhD students:
  Hyotae Kim, Xiaotian Zheng, Chunyi Zhao, Yunzhe Li, Jizhou Kang, Zach Horton

- PhD alumni:
  Matt Heiner, Yifei Yan, Annalisa Cadonna, Robert Richardson, Sai Xiao, Maria DeYoreo, Valerie Poynor, Ziwei Wang, Marian Farah, Kassandra Fronczyk, Matt Taddy, Milovan Krnjajić

- UCSC colleagues:
  Bruno Sansó, Abel Rodriguez, Juhee Lee, Raquel Prado, Steve Munch

- Funding from NSF under awards SES 1950902 and DMS 2015428