Nonparametric Bayesian density regression: modeling methods and applications

Athanasios Kottas

Department of Statistics, University of California, Santa Cruz

Bayesian Nonparametrics Networking Workshop 2023
Monash University, Caulfield campus, Melbourne, Australia
December 4-7, 2023

Outline

1. Bayesian nonparametrics: introduction and motivation
2. Dirichlet process; DP mixture models; Dependent Dirichlet processes
3. Nonparametric Bayesian density regression
4. Fully nonparametric quantile regression
5. Density regression with ordinal responses
6. Density regression in survival analysis

Bayesian nonparametrics: introduction and motivation

Parametric vs. nonparametric Bayes: a simple example

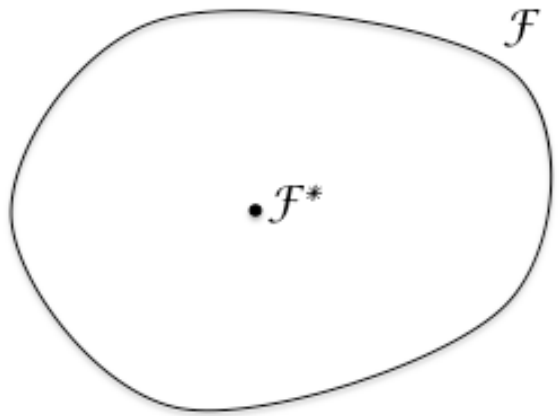
- Let $y_i \mid G \overset{i.i.d.}{\sim} G$, with $G \in \mathcal{F}^*$,

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

- 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 **random distributions** (or **random 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.

- 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

- An 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 or LAD are classical semiparametric estimation techniques: they estimate the regression coefficients $\boldsymbol{\beta}$ *without* assuming a probability model for the error distribution.

- In contrast, a Bayesian semiparametric modeling approach proceeds 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

- Parametric modeling: based on parametric families of distributions $\{G(\cdot;\theta) : \theta \in \Theta\} \rightarrow$ requires prior distributions over Θ.
- Seek a richer class, i.e., $\{G : G \in \mathcal{G}\} \rightarrow$ requires *nonparametric* prior distributions over $\mathcal{G}$.
- How to choose $\mathcal{G}$? how to specify the prior over $\mathcal{G}$? $\rightarrow$ requires specifying prior distributions for infinite-dimensional parameters.

- What makes a nonparametric model "good"? (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.

Methods for construction of NPB models for distributions

- The object is to define priors over spaces of distributions G 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

The Dirichlet process

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.

- For uncountable spaces, such as $\mathcal{X} = \mathbb{R}$, consider finite collections of (measurable) subsets of $\mathcal{X}$, say, $B_1, ..., B_k$, that form a partition of $\mathcal{X}$.
  - The Dirichlet distribution is a natural candidate for the distribution of the probability vector $(G(B_1), ..., G(B_k))$.
  - But care is needed, a system of Dirichlet f.d.d.s must be consistent with *any* other partition (any 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 of the Dirichlet process

- The DP is characterized by two parameters:
  - $\alpha \rightarrow$ a positive scalar parameter;
  - $G_0 \rightarrow$ a specified probability measure (distribution) on $\mathcal{X}$.

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

$$(G(B_1), ..., G(B_k)) \sim \text{Dirichlet}(\alpha G_0(B_1), ..., \alpha G_0(B_k)).$$

  - Here, $G(B_i)$ (a random variable) and $G_0(B_i)$ (a constant) denote the probability of set B_i under G and G_0, respectively.

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 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 $(G(B_1), ..., G(B_k))$ in the DP definition.

Interpreting the parameters of the Dirichlet process

- For any measurable subset B of $\mathcal{X}$, we have from the definition that $G(B) \sim \text{Beta}(\alpha G_0(B), \alpha G_0(B^c))$, and thus

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

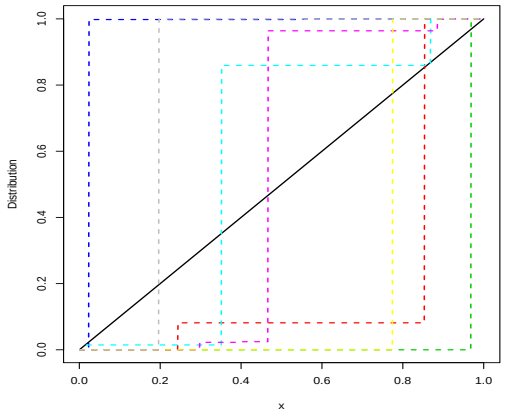
- G_0 plays the role of the *center* of the DP (also referred to as the baseline distribution).

- 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 G from the process to be to G_0.

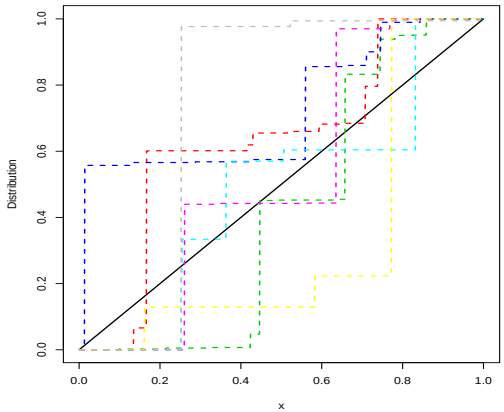
- Chapter 4 of Ghosal and van der Vaart (2017) provides a detailed account of several properties of the Dirichlet process.

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

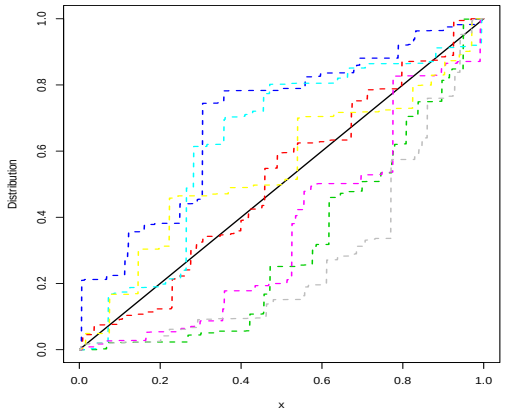
$\alpha = 0.1$



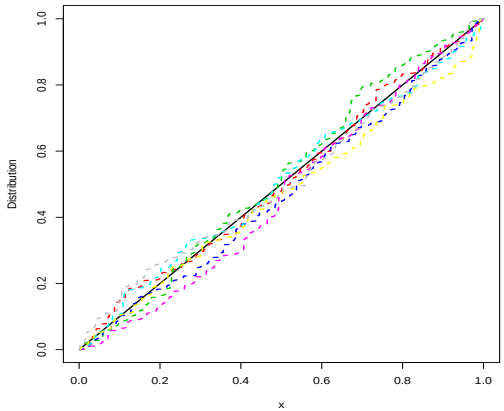
$\alpha = 1$



$\alpha = 10$



$\alpha = 100$



DP($\alpha, G_0 = \text{Unif}(0,1)$) c.d.f. realizations. The solid black line corresponds to the baseline c.d.f., while the dashed colored lines represent multiple realizations.

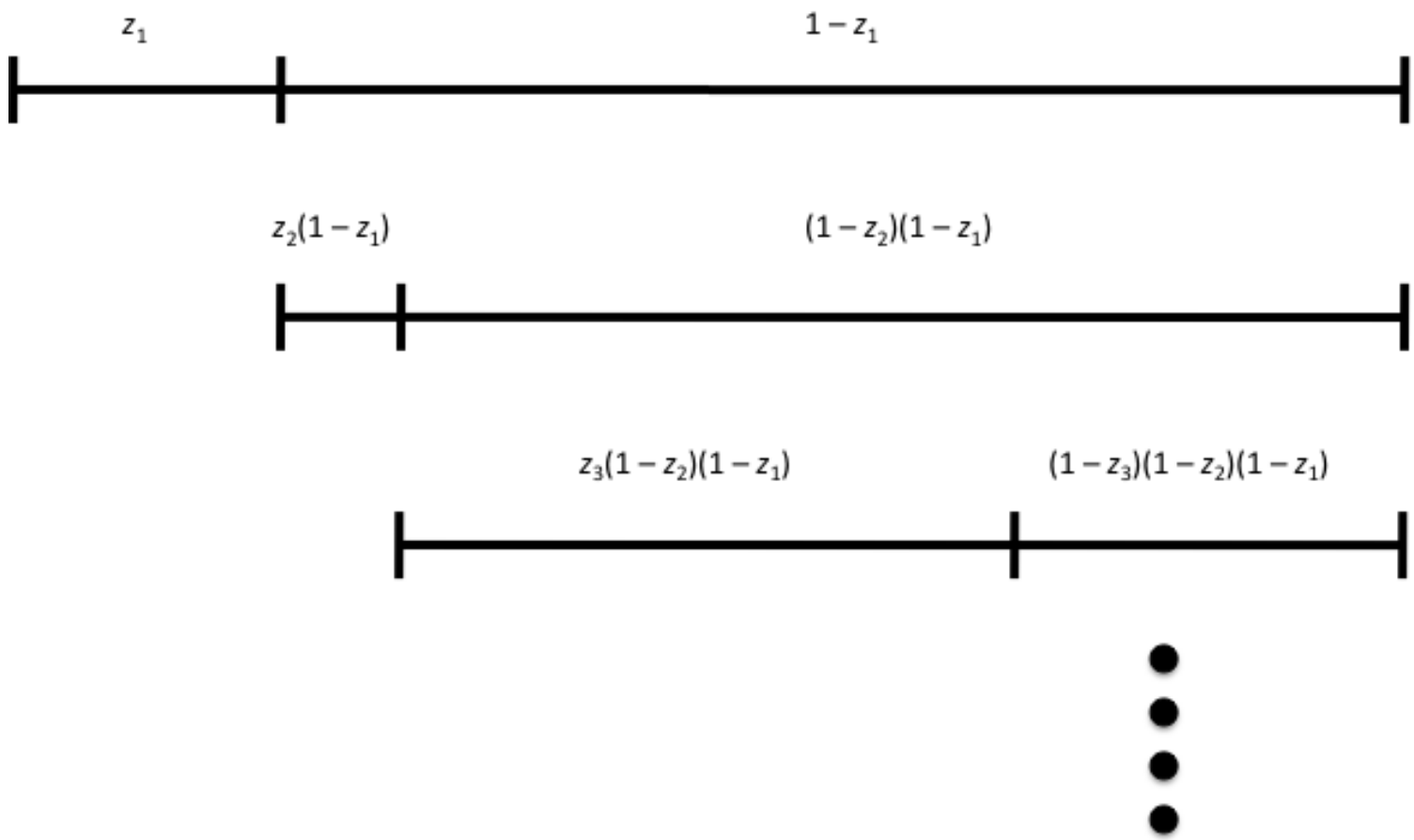
Constructive definition of the DP

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

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

  where δ_a denotes a point mass at a.
- Hence, the DP generates discrete distributions (proved earlier by Ferguson, 1973, and Blackwell, 1973).

The stick-breaking construction



- 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 $\rightarrow$ in fact, it provides (up to a truncation approximation) the entire distribution G, not just c.d.f. sample paths.
- For example, a possible approximation is $G_J = \sum_{j=1}^{J} p_j \delta_{\vartheta_j}$, with $p_j = \omega_j$ for $j = 1, ..., J-1$, and $p_J = 1 - \sum_{j=1}^{J-1} \omega_j = \prod_{r=1}^{J-1}(1 - z_r)$.
- To specify J, a simple approach involves working with the expectation for the partial sum of the stick-breaking weights:

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

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

More on the constructive definition of the DP

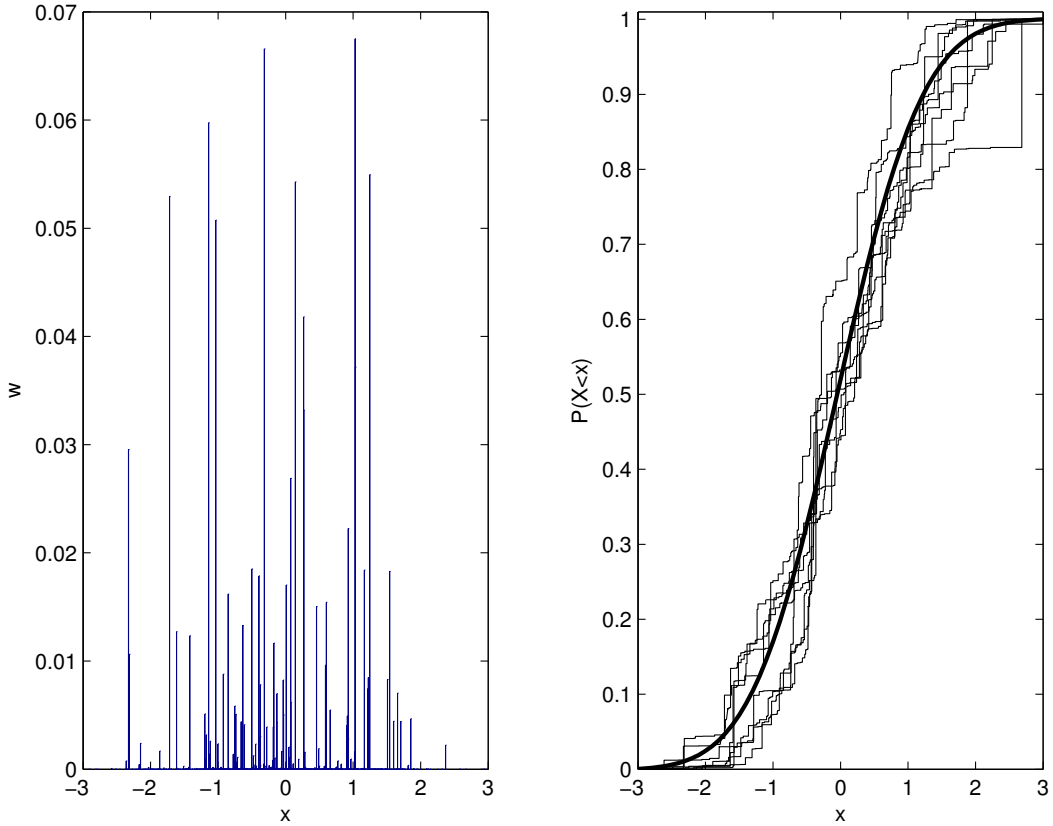


Illustration for a DP with $G_0 = \mathrm{N}(0, 1)$ and $\alpha = 20$. In the left panel, the spiked lines are located at 1000 $\mathrm{N}(0, 1)$ draws with heights given by the (truncated) stick-breaking weights. These spikes are then summed to generate one c.d.f. sample path. The right panel shows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

Generalizing the DP

Many random probability measures can be defined by means of a stick-breaking construction $\rightarrow$ the z_r are drawn independently from a distribution on $[0, 1]$.

- For example, the Beta two-parameter process (Ishwaran and Zarepour, 2000) is defined by choosing $z_r \sim \text{Beta}(a, b)$.
- If $z_r \sim \text{Beta}(1 - a, b + ra)$, for $r = 1, 2, \ldots$ and some $a \in [0, 1)$ and $b \in (-a, \infty)$ we obtain the two-parameter Poisson-Dirichlet process (e.g., Pitman and Yor, 1997).
- The general case, $z_r \sim \text{Beta}(a_r, b_r)$ (Ishwaran and James, 2001).
- Probit stick-breaking: $z_r = \Phi(x_r)$, where $x_r \sim N(\mu, \sigma^2)$ and Φ is the standard normal c.d.f. (Rodríguez and Dunson, 2011).
- Logit stick-breaking: $z_r = \exp(x_r)/\{1 + \exp(x_r)\}$, where $x_r \sim N(\mu, \sigma^2)$ (Rigon and Durante, 2021).

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 $G \sim \text{DP}(\alpha, G_0)$. Then, the posterior distribution of G is the $\text{DP}(\tilde{\alpha}, \tilde{G}_0)$, where

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

- 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

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

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

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.

The DP prediction rule (Pólya urn scheme)

- Start with $X_i \mid G \overset{i.i.d.}{\sim} G$, for $i = 1, ..., n$, and $G \sim \text{DP}(\alpha, G_0)$.
  - 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,

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

  - 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 $G \sim \text{DP}(\alpha, G_0)$ implies the marginal $\Pr(X \in B) = \int \Pr(X \in B \mid G)\, \text{d}\mathcal{P}(G) = \int G(B)\, \text{d}\mathcal{P}(G) = \text{E}(G(B)) = G_0(B)$.

Pólya urn characterization of the DP

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

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

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

Pólya urn characterization of the DP

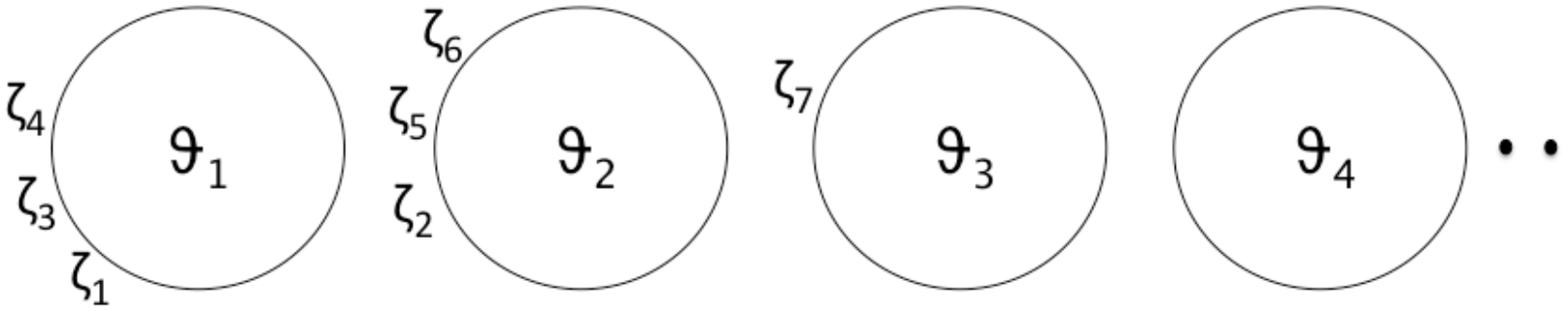
- The forward direction is readily obtained from the DP prediction rule.
- Blackwell and MacQueen (1973) proved the other direction, thus, characterizing the DP as the de Finetti measure for *Pólya sequences*.
- A sequence of r.v.s, $\{X_n : n \geq 1\}$, (w.l.o.g. on $\mathbb{R}$) is a Pólya sequence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

The Chinese restaurant process

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

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

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



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 DP(α, G_0) is the de Finetti measure for this exchangeable sequence.

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 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.
- Strong connections with finite mixture 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 can be implemented 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 $i = 1, \ldots, n$, 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) $\rightarrow$ 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**

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

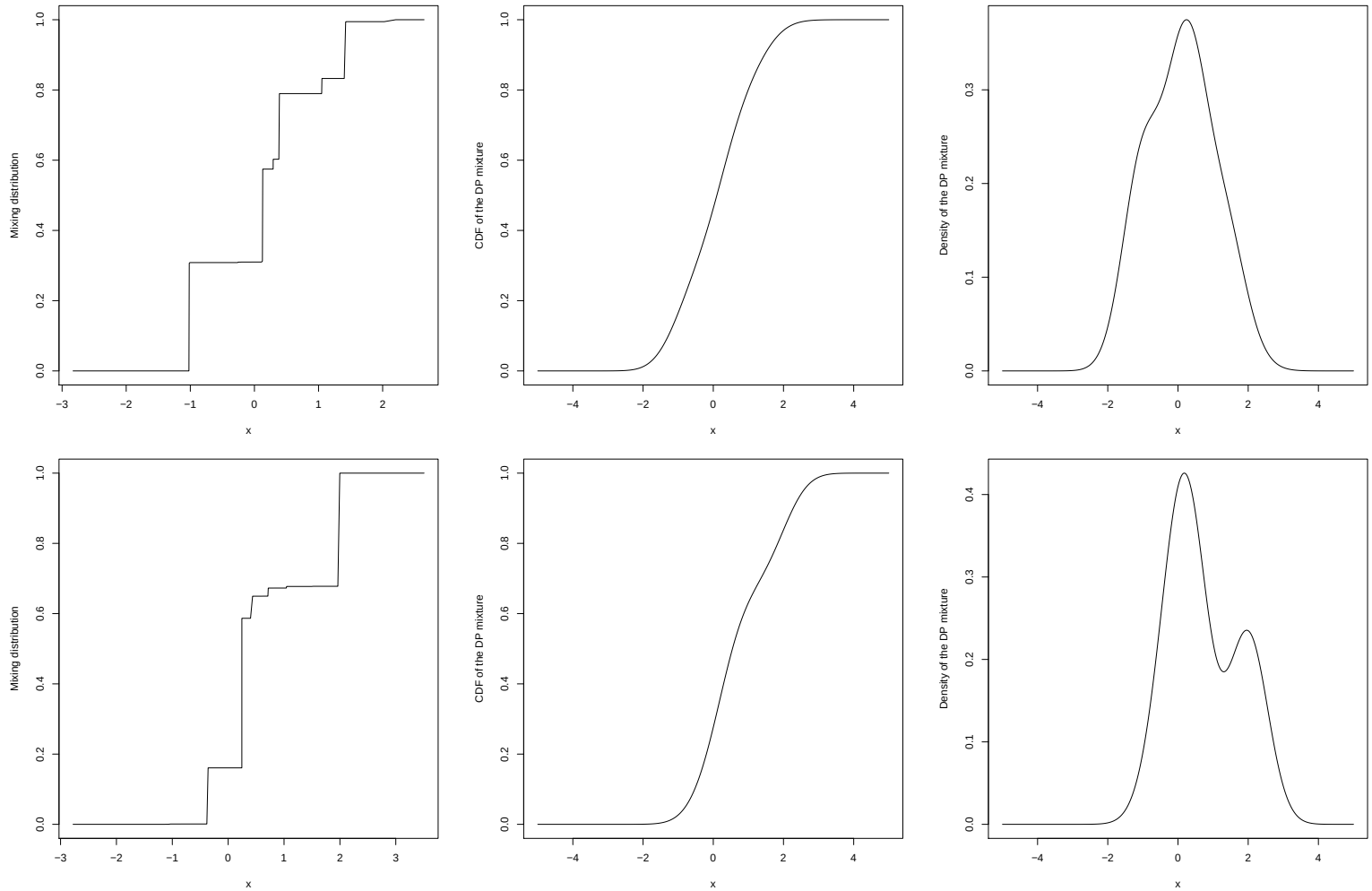
where $K(y \mid \theta)$ is a parametric c.d.f. (with parameters θ).

- 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 $K(y \mid \theta)$.



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 $\rightarrow$ 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 (atoms): ϑ_ℓ i.i.d. G_0

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).

- 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,

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

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

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

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

Semiparametric Dirichlet process mixture models

- In many applications, semiparametric DP mixtures are employed

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

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

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 $\boldsymbol{x}_i$

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

  - 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:

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

- Alternatively, with discrete latent variables L_i:

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

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 $\alpha \to 0^+$

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

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

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

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

Prior specification

- Taking expectation over G with respect to its DP prior DP(α, G_0),

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

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

- On the other hand, recall that for the DP(α, 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).

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.

Number of distinct components

- A fairly accurate approximation:

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

  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:

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

  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).

Dependent Dirichlet processes

Nonparametric priors for dependent distributions

- In many applications, the objective is to model a collection of distributions $\mathcal{G} = \{G_s : s \in S\}$, indexed by $s \in S$
  - 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., $G_s \equiv G \sim \text{DP}(\alpha, G_0)$ for all s. This is too restrictive.
  - Assume that the distributions are independent and identically distributed, e.g., $G_s \sim \text{DP}(\alpha, G_0)$ independently for each s. This is wasteful.
- We would like something in between.

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, e.g., product of mixtures of DPs (Cifarelli and Regazzini, 1978). Simple but restrictive.
  - Priors for a finite number of distributions through linear combinations of realizations from independent DPs (Müller et al., 2004).
  - Hierarchical nonparametric priors for finite collections of distributions: analysis of densities model (Tomlinson and Escobar, 1999); hierarchical DP (Teh. et al., 2006); nested DP (Rodriguez et al., 2008).
  - 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).

Definition of the dependent Dirichlet process

- Recall the DP constructive definition: if $G \sim \text{DP}(\alpha, G_0)$, 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. Beta$(1, \alpha)$.

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

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

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

Dependent Dirichlet processes

- For any fixed s, this construction yields a DP prior for distribution G_s.
- 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 $G_{s'}$ driven by the distance between spatial sites s and s$'$.
- For specified set A, $\{G_s(A) : s \in S\}$ is a stochastic process with beta marginals. The covariance between $G_s(A)$ and $G_{s'}(A)$ 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" DDP models: the weights do not depend on s; dependence is induced across atoms in the stick-breaking construction:

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

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

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

"Common-atoms" dependent Dirichlet processes

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

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

where the θ_ℓ are i.i.d. from G_0.

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

Nonparametric Bayesian density regression

Density regression using Dirichlet process mixtures

- Dominant trend in the Bayesian regression literature: seek flexible regression function models, and accompany them 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.

- 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.

Density regression using Dirichlet process mixtures

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

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

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

- 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

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

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

- The implied conditional response density is a normal mixture with covariate-dependent mixture weights:

$$f(y \mid \boldsymbol{x}, G) = \sum_{\ell=1}^{\infty} q_\ell(\boldsymbol{x}) \, \mathrm{N}(y \mid \lambda_\ell(\boldsymbol{x}), \tau_\ell^2)$$

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 $\tau_\ell^2 = \Sigma_\ell^y - \Sigma_\ell^{y\boldsymbol{x}} (\Sigma_\ell^{\boldsymbol{x}})^{-1} (\Sigma_\ell^{y\boldsymbol{x}})^T$
  - using 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.

Mean regression functional under normal DP mixtures

- Mean regression function:

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

  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 $r = 1, \ldots, p$, are the elements of vector $\Sigma_\ell^{y\boldsymbol{x}} (\Sigma_\ell^{\boldsymbol{x}})^{-1}$

- The density regression approach (under a normal mixture for the joint response-covariate distribution) implies a mixture of linear regressions for the mean regression function, with covariate-dependent mixture weights.

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 $n = 2000$.

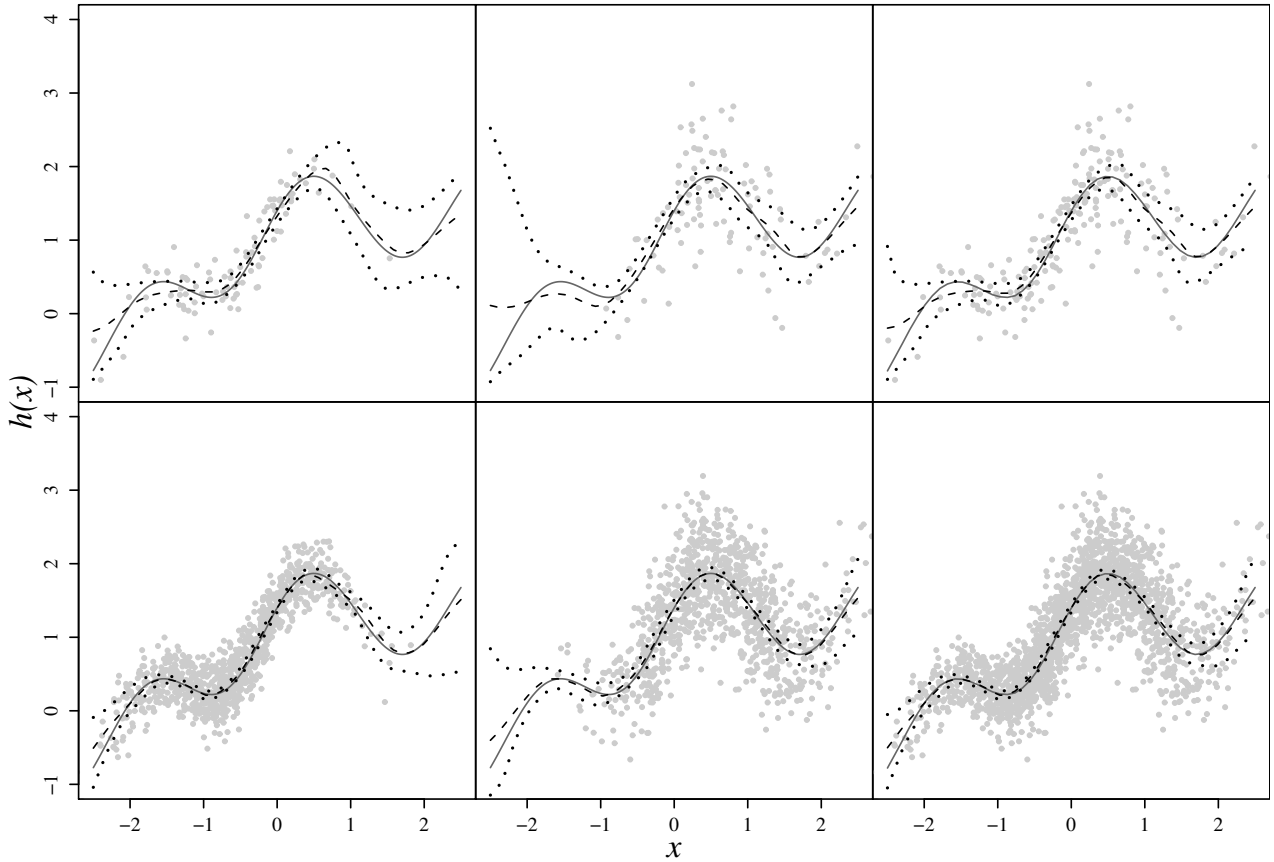
- DP mixture model with a mixed normal/Bernoulli kernel:

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

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



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

DP mixture density regression: applications

- Regression modeling with categorical responses (Shahbaba and Neal, 2009; Dunson and Bhattacharya, 2011; Hannah et al., 2011; DeYoreo and Kottas, 2015, 2018a,b).
- Functional data analysis through density estimation (Rodriguez et al., 2009).
- Fully nonparametric quantile regression (Taddy and Kottas, 2010).
- Product partition models with regression on covariates (Müller and Quintana, 2010; Park and Dunson, 2010), and regression modeling with *enriched* DP priors (Wade et al., 2014).
- Inference for marked Poisson processes (Taddy and Kottas, 2012; Xiao et al., 2015).
- Nonparametric survival regression (Poynor and Kottas, 2017).
- Density autoregression, including lag selection (Heiner and Kottas, 2022).

Fully nonparametric quantile regression

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 p-variate vector of (continuous) covariates $\boldsymbol{x}$ with a DP mixture of normals:

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

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

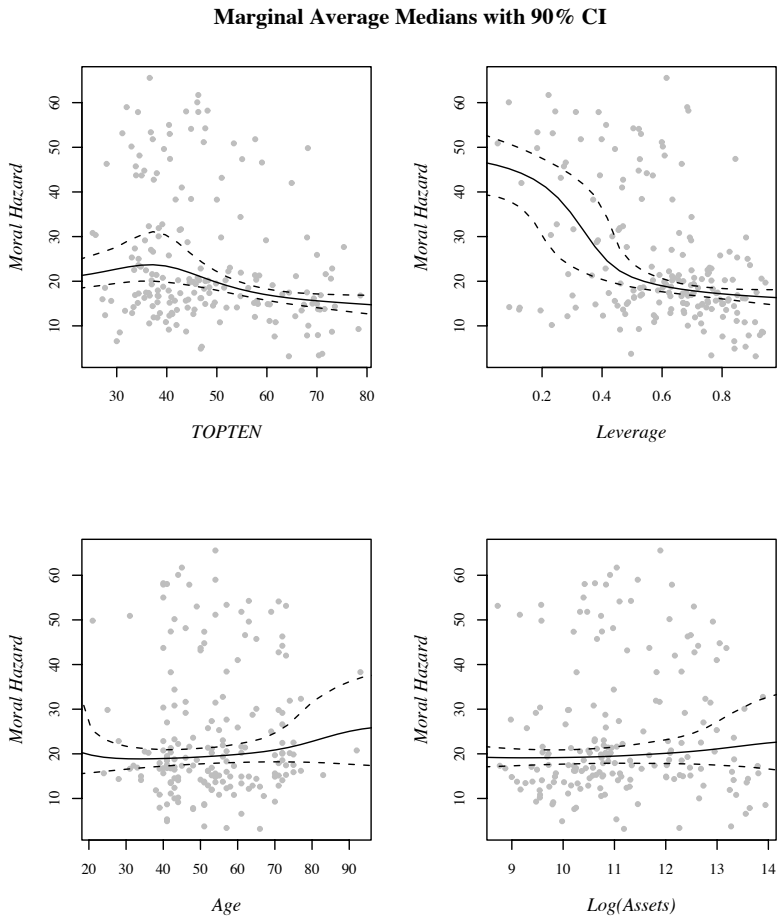
- For any grid of values $(y_0, \boldsymbol{x}_0)$, obtain posterior samples for:
  - Conditional density $f(y_0 \mid \boldsymbol{x}_0, G)$ and conditional c.d.f. $F(y_0 \mid \boldsymbol{x}_0, G)$.
  - Conditional quantile regression $q_p(\boldsymbol{x}_0 \mid G)$, for any $0 < p < 1$.

- 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

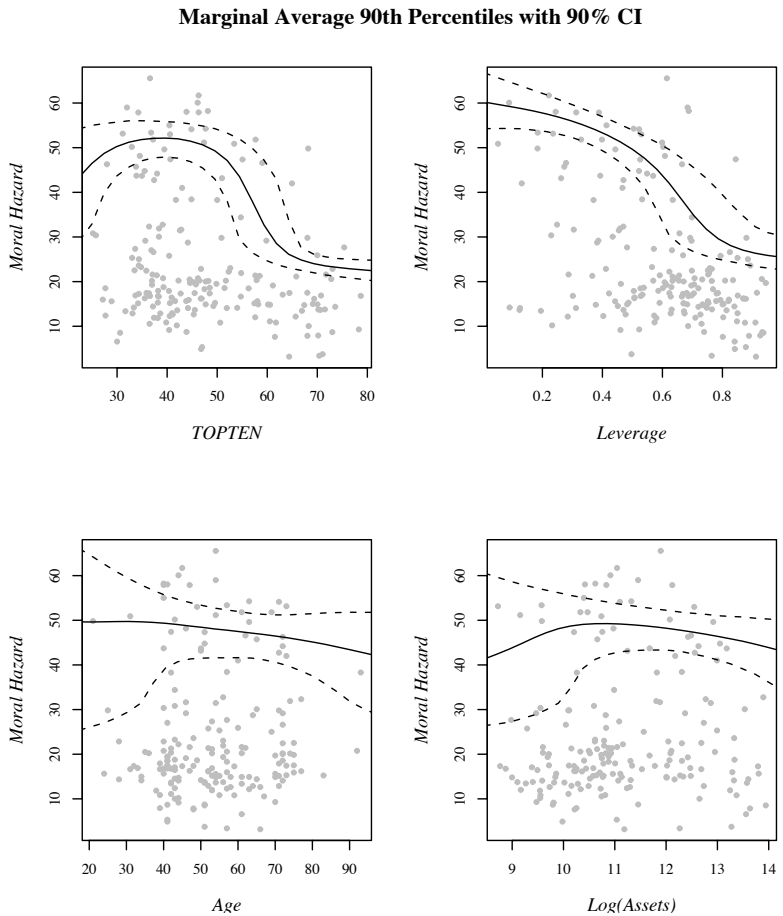
- *Moral hazard* data on the relationship between shareholder concentration and several indices for managerial moral hazard in the form of expenditure with scope for private benefit (Yafeh & Yoshua, 2003).
  - Data set includes a variety of variables describing 185 Japanese industrial chemical firms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



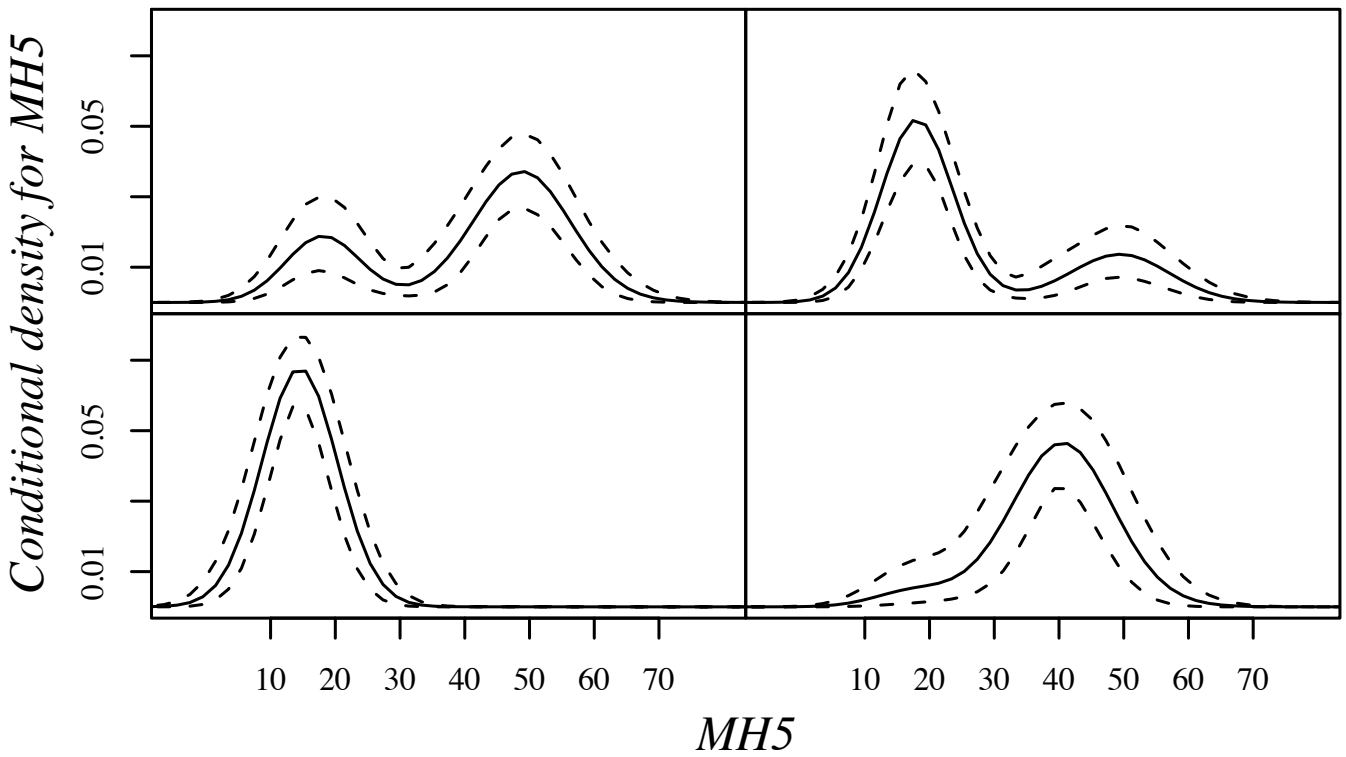
Posterior mean and 90% interval estimates for median regression for *MH*5 conditional on each individual covariate. Data scatterplots are shown in grey.

Quantile regression: data example



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

Quantile regression: data example



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*))

Density regression with ordinal responses

Density regression for ordinal responses

- Assume each ordinal response represents a discretized version of a latent continuous response.
- k ordinal variables $\boldsymbol{Y} = (Y_1, \ldots, Y_k)$, with $y_j \in \{1, \ldots, C_j\}$, and p (continuous) covariates $\boldsymbol{X} = (X_1, \ldots, X_p)$.
- Assume

$$Y_j = \ell \quad \text{if-f} \quad \gamma_{j,\ell-1} < Z_j \leq \gamma_{j,\ell}, \qquad j = 1, ..., k; \;\; \ell = 1, ..., C_j$$

  (with $\gamma_{j,0} = -\infty$ and $\gamma_{j,C_j} = \infty$).

- Multivariate normal distribution for $\boldsymbol{Z} = (Z_1, \ldots, Z_k) \rightarrow$ multivariate ordinal probit model.
  - Symmetric, unimodal latent response distribution with mean $\boldsymbol{x}^T\beta \rightarrow$ implies restrictive covariate effects on probability response curves.
  - Computational challenges in estimating cut-off points.

Density regression for ordinal responses

- Model the joint distribution of the latent continuous responses, $\boldsymbol{Z}$, and the covariates, $\boldsymbol{X}$, with a multivariate normal DP mixture:

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

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

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

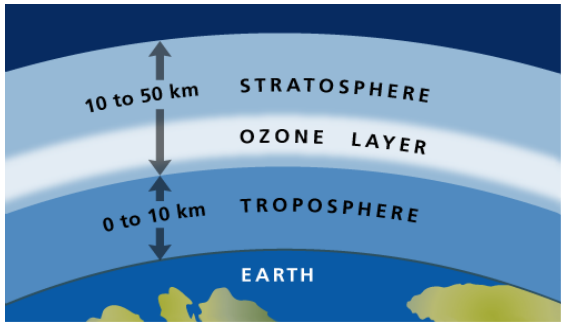
  - with covariate-dependent weights $w_r(\boldsymbol{x}) \propto p_r\, \mathrm{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_r^x, \Sigma_r^{xx})$
  - and covariate-dependent probabilities, where $m_r(\boldsymbol{x}) = \boldsymbol{\mu}_r^z + \Sigma_r^{zx}(\Sigma_r^{xx})^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_r^x)$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

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).

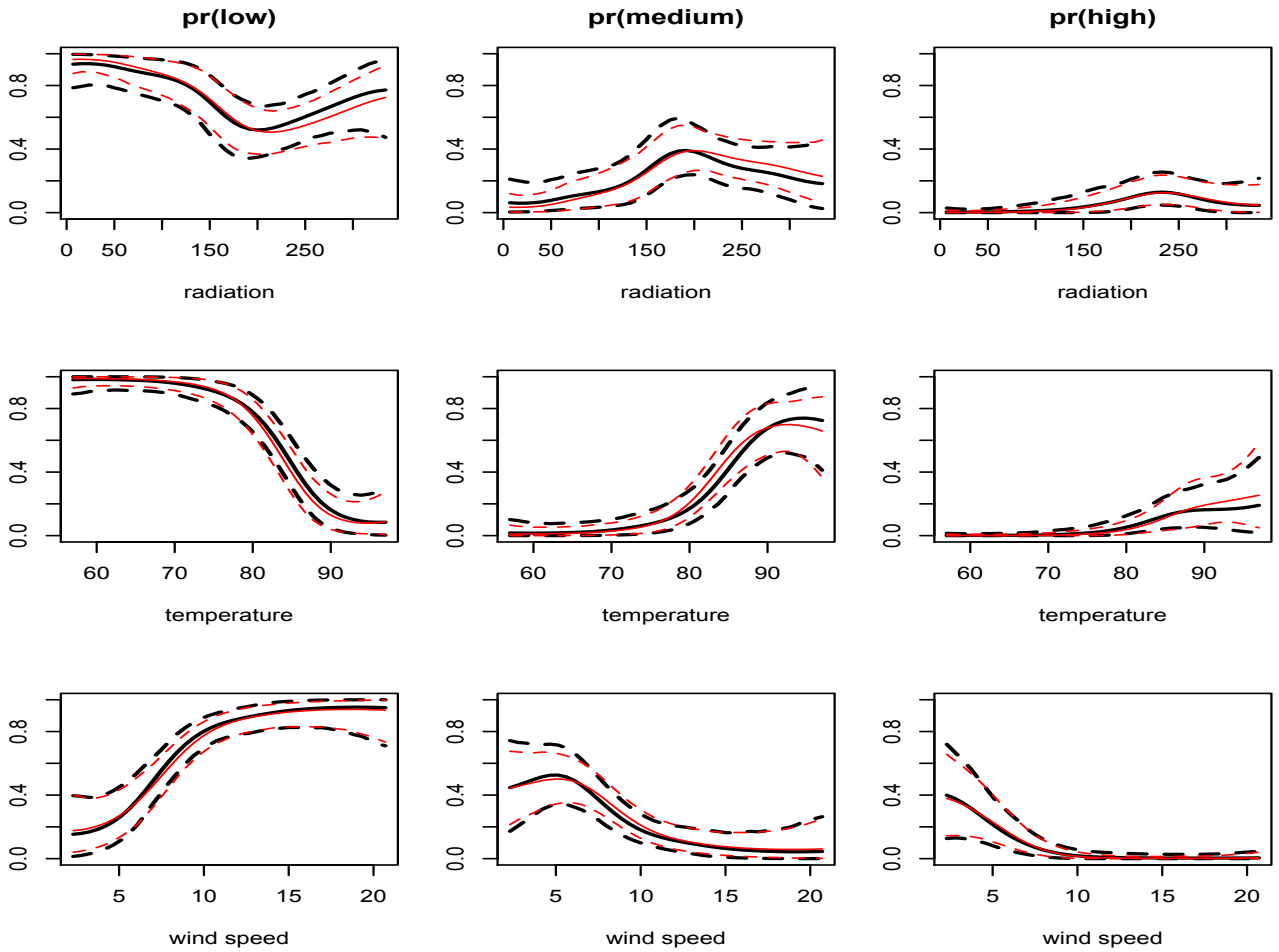
Ozone concentration data example

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



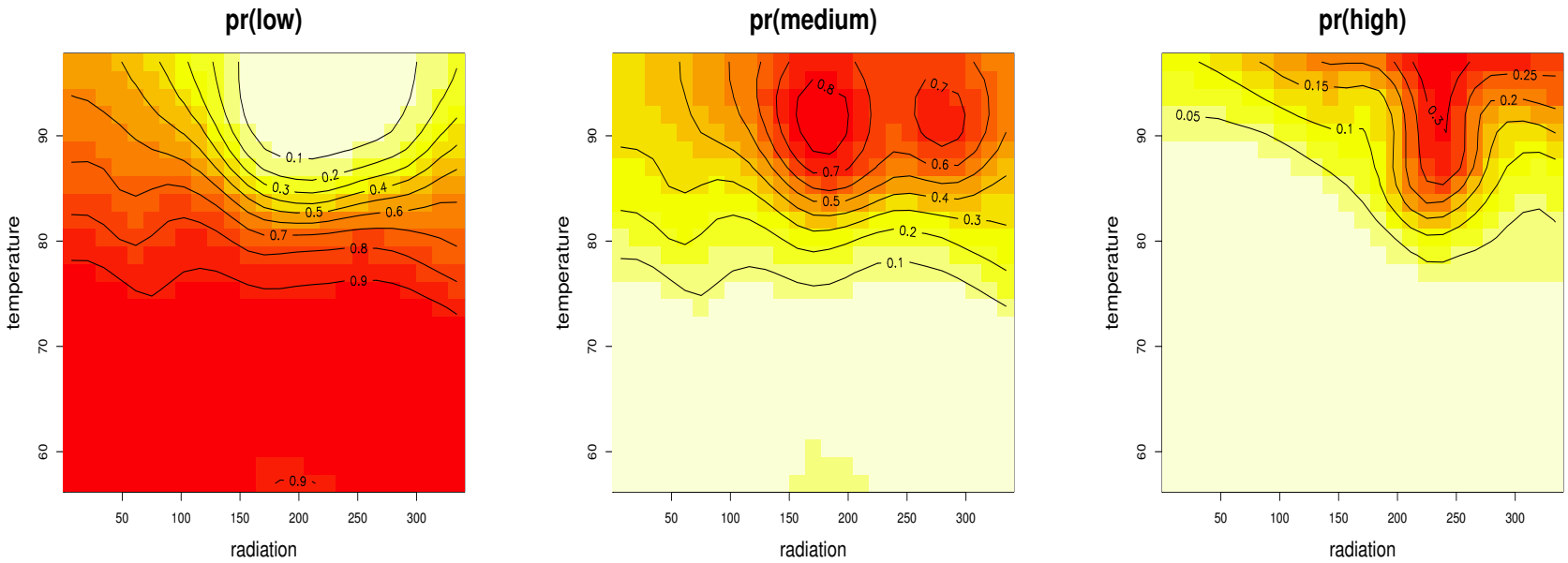
- Ozone concentration recorded on continuous scale.
- To construct an ordinal response: define "high" as above 100 ppb, "medium" as $(50, 100]$ ppb, and "low" as less than 50 ppb.
- Comparison of inferences from the model for $(Y, \boldsymbol{X})$ with those from a DP mixture of normals model for $(Z,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

Ozone concentration data example



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

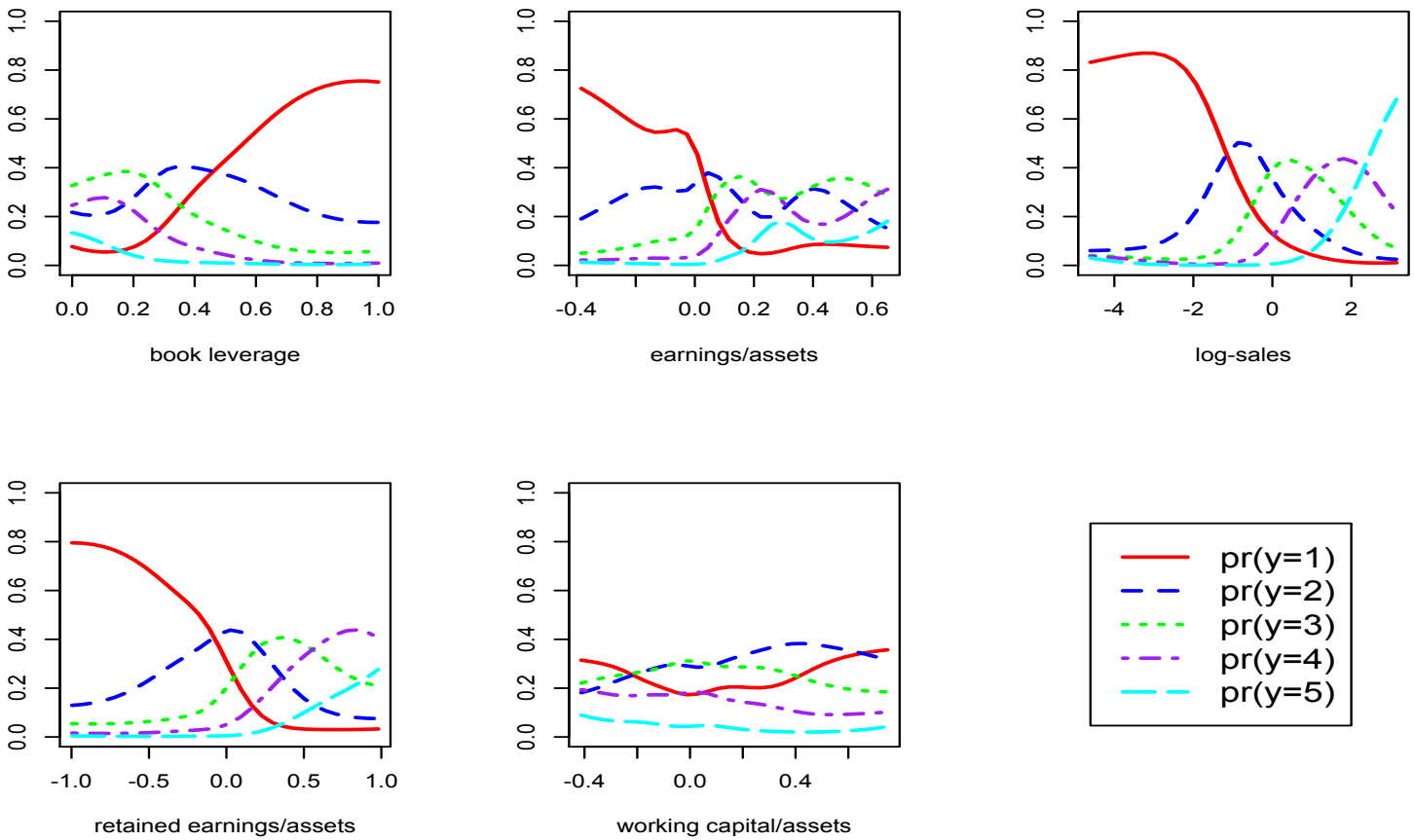


Posterior mean estimates for $\Pr(Y = \ell \mid x_1, x_2, G)$, for $\ell = 1, 2, 3$, corresponding to low (left), medium (middle) and high (right). Red represents a value of 1, white represents 0.

Credit ratings of U.S. companies

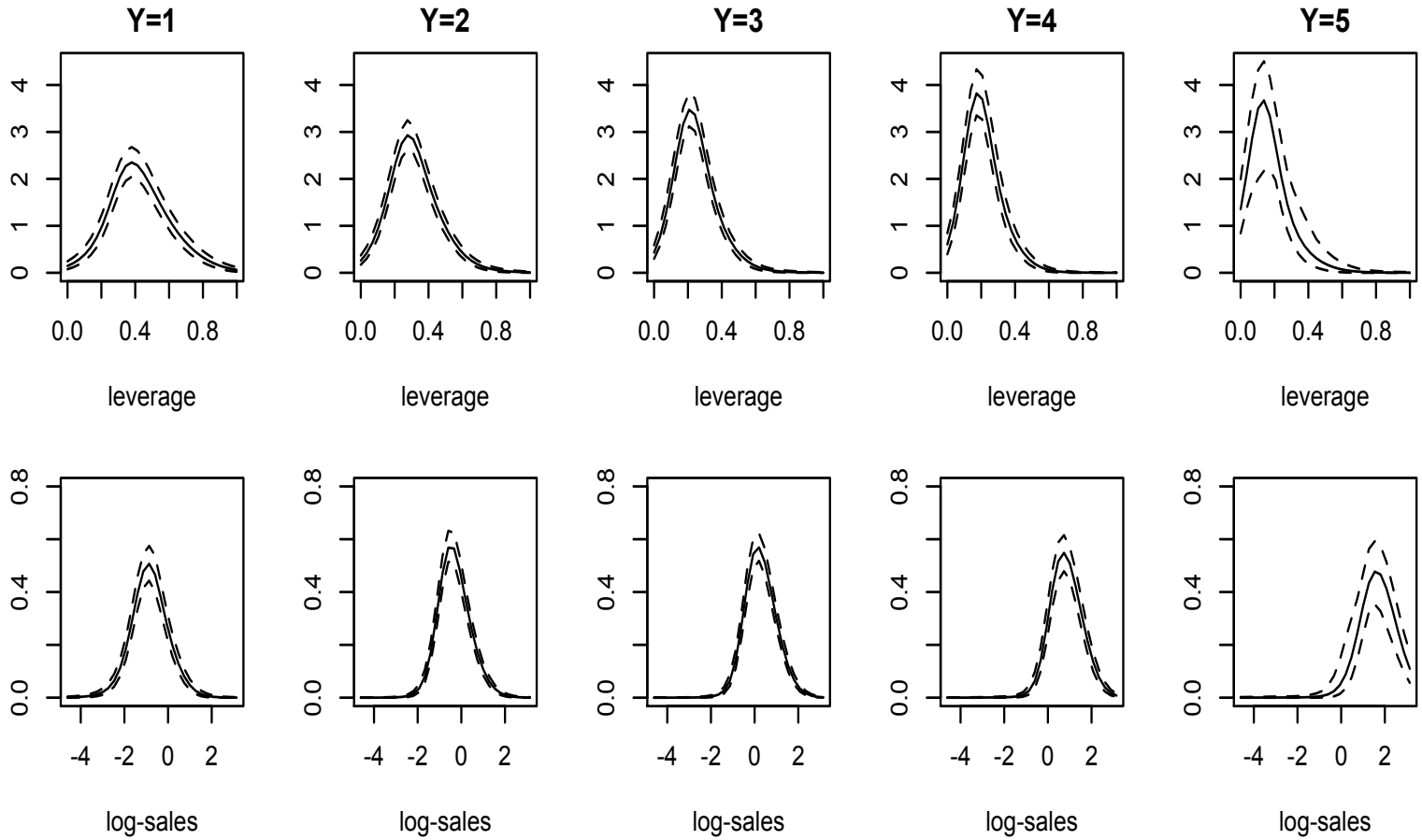
- Data on Standard and Poor's (S&P) credit ratings for 921 U.S. firms in year 2005 (Verbeek, 2008; Chib & Greenberg, 2010).
- Credit rating recorded on a five-point ordinal scale, where higher ratings indicate more creditworthiness
- Five covariates (firm characteristics)
  - X_1: book leverage (ratio of debt to assets)
  - X_2: earnings before interest and taxes / total assets
  - X_3: standardized log-sales (proxy for firm size)
  - X_4: retained earnings / total assets (proxy for historical profitability)
  - X_5: working capital / total assets (proxy for short-term liquidity)

Credit ratings of U.S. companies



Posterior mean estimates for $\Pr(Y = l \mid x_m, G)$, for each covariate $m = 1, \ldots, 5$.

Credit ratings of U.S. companies



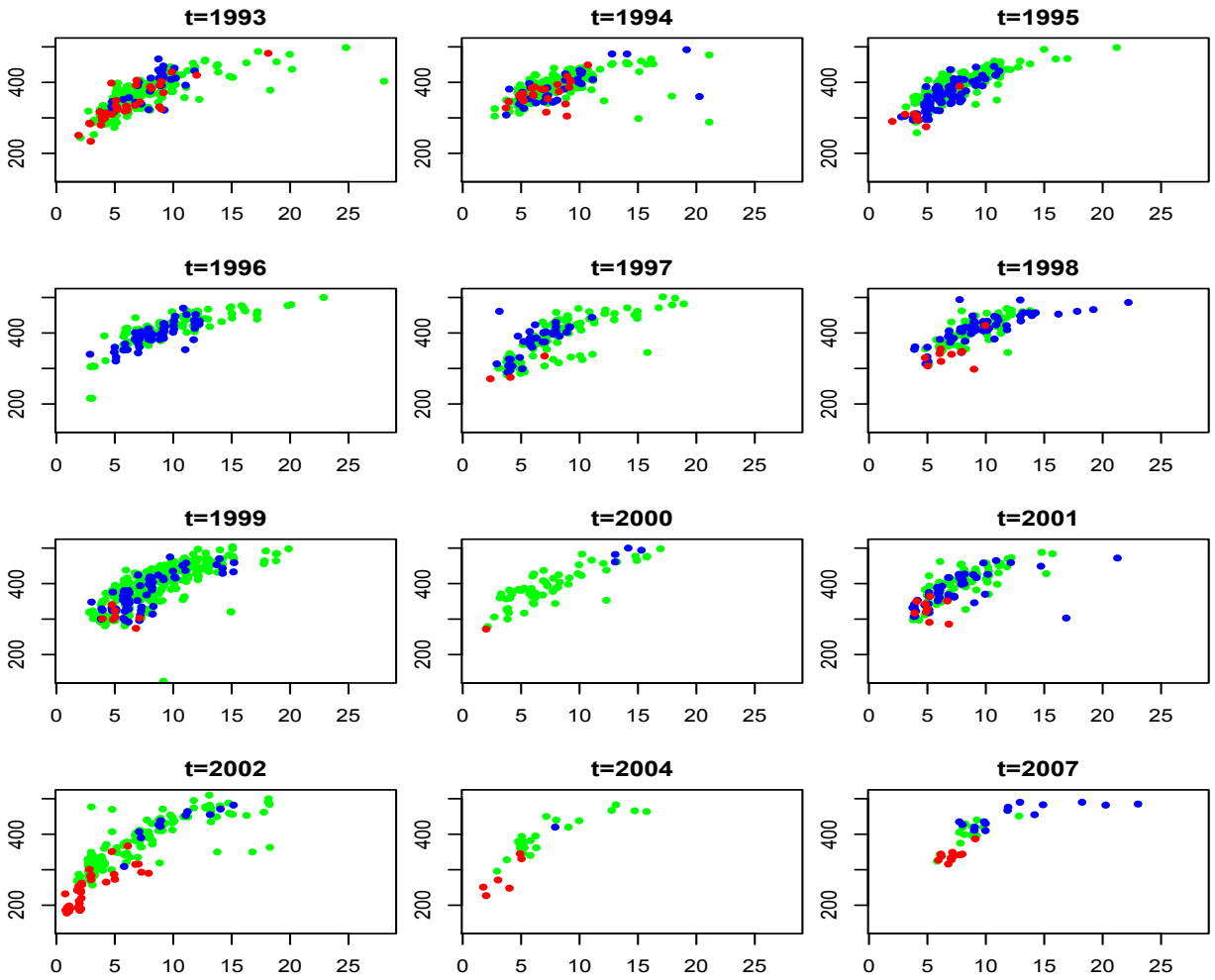
Posterior mean and 95% interval estimates for covariate densities $f(x \mid Y = l, G)$ conditional on ordinal credit rating $l = 1, ..., 5$. The top row corresponds to covariate book leverage, and the bottom row to standardized log-sales.

Extension to dynamic ordinal regression modeling

- Focusing on a univariate ordinal response, we seek to extend to a model for $\Pr_t(Y \mid \boldsymbol{x})$, for $t \in \mathcal{T} = \{1, 2, \dots\}$
- Build on the earlier framework by extending to a prior model for $\{f(z, \boldsymbol{x} \mid G_t) : t \in \mathcal{T}\}$, and thus for $\{\Pr(Y \mid \boldsymbol{x}, G_t) : t \in \mathcal{T}\}$

- 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) $\rightarrow$ incorporate age into the model in the same fashion with the maturity variable.

Rockfish data



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:

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

- 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:

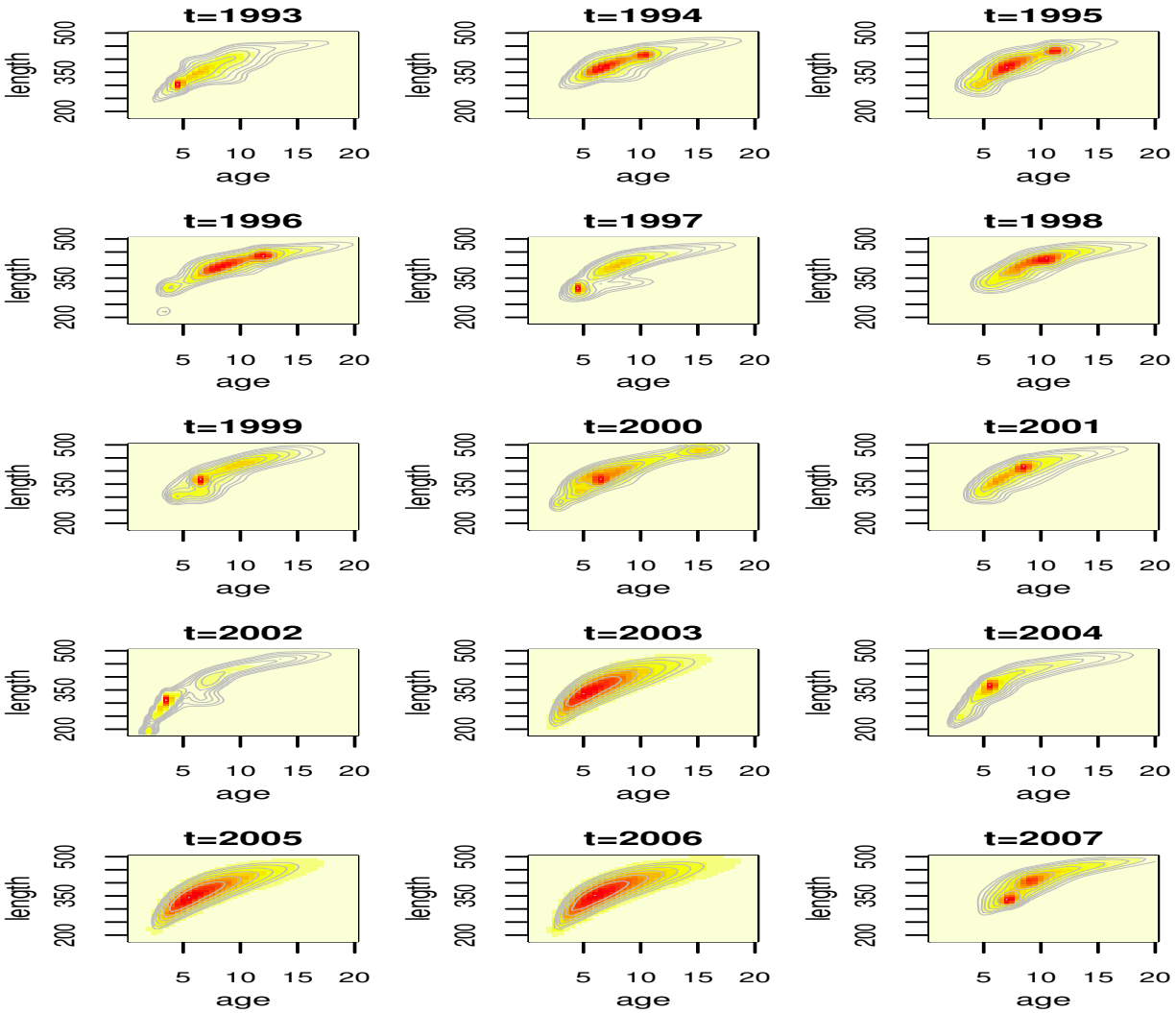
$$\mathcal{B} = \left\{ \beta_t = \exp\left(-\frac{\zeta^2 + \eta_t^2}{2\alpha} \right) : t \in \mathcal{T} \right\}$$

  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 $\eta_{r,1} \overset{ind.}{\sim} \mathrm{N}(0,1)$)

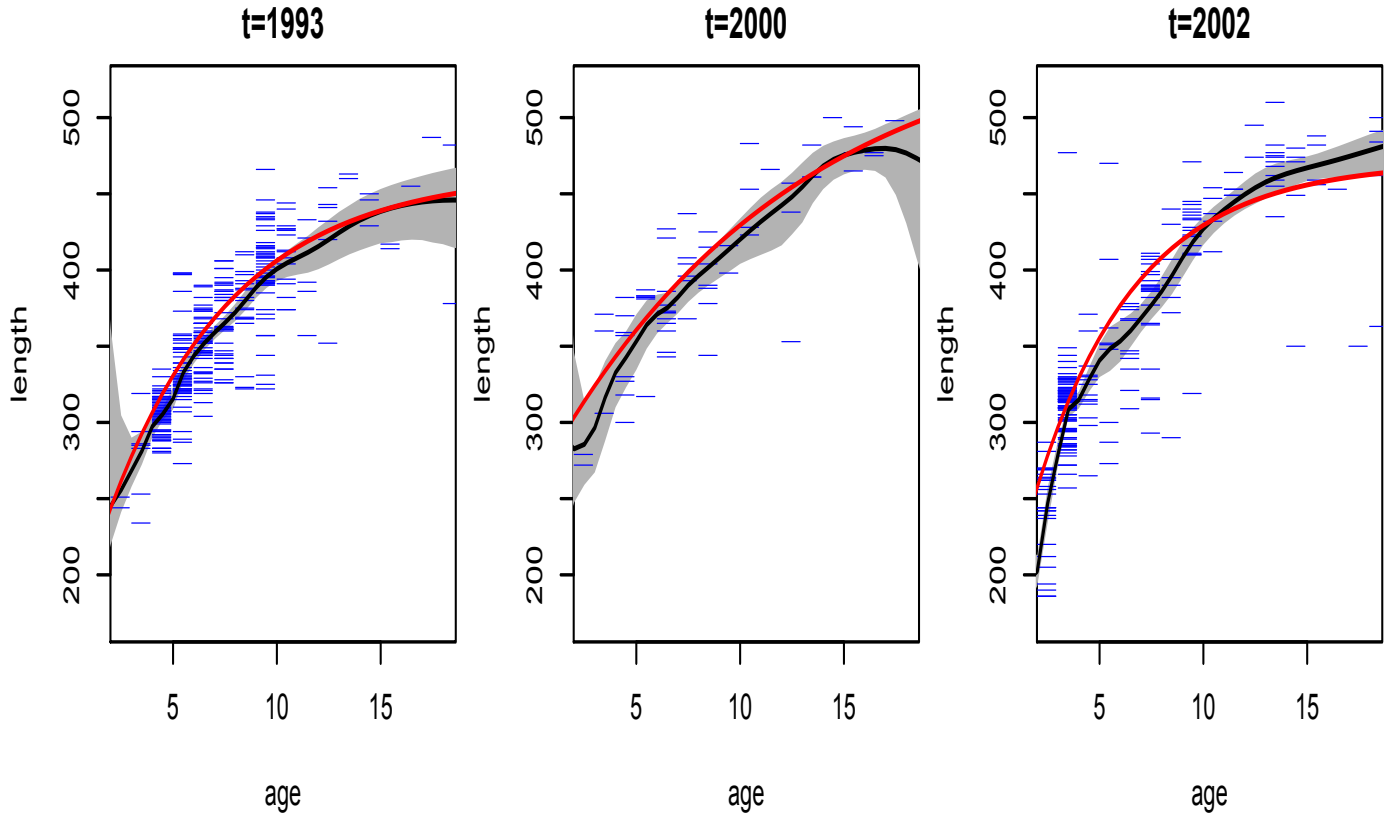
- Different types of correlations can be studied, e.g., corr$(G_t(A), G_{t+1}(A))$, for any subset A in the support of the G_t .

Rockfish data



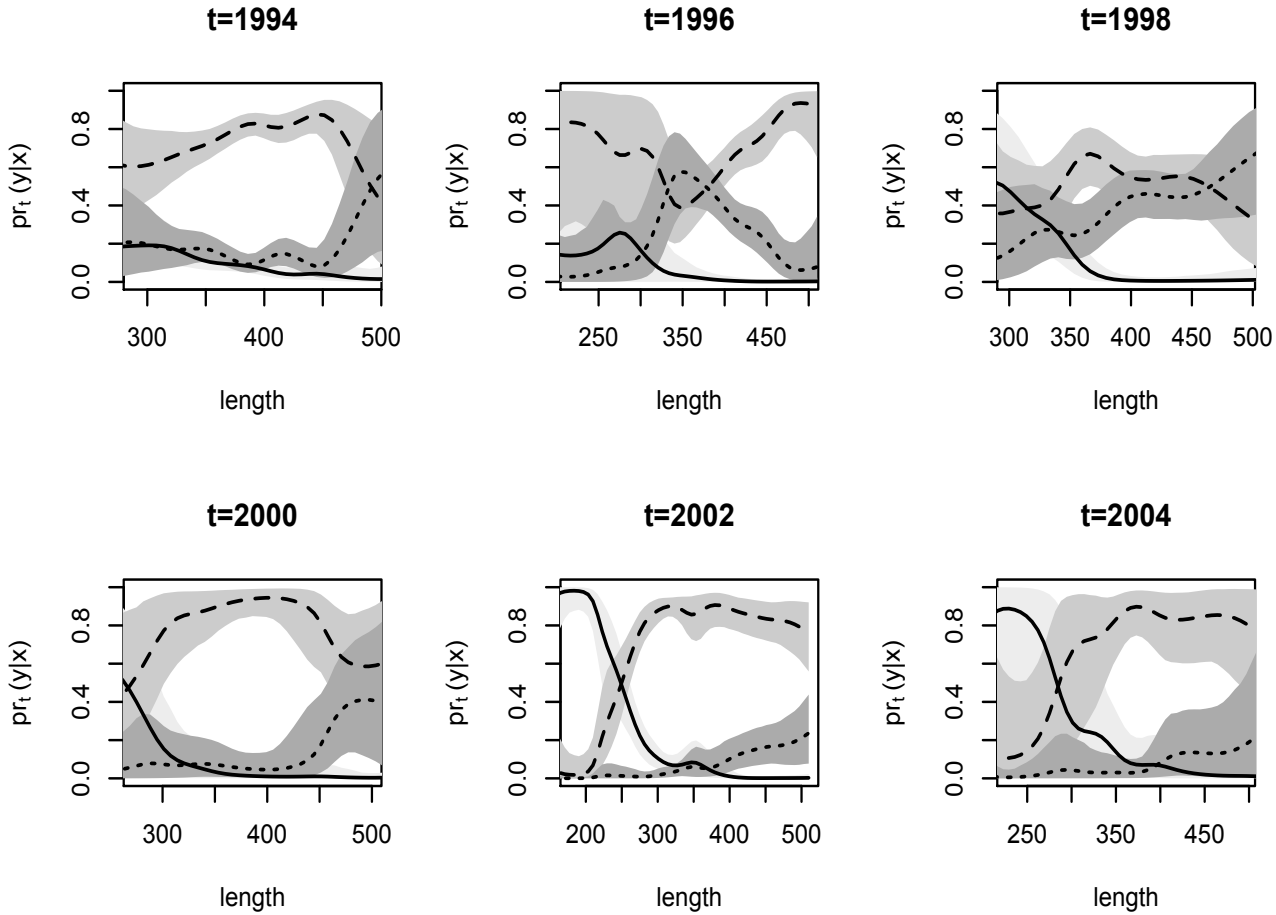
Posterior mean estimates for $f(\text{age}, \text{length})$.

Rockfish data



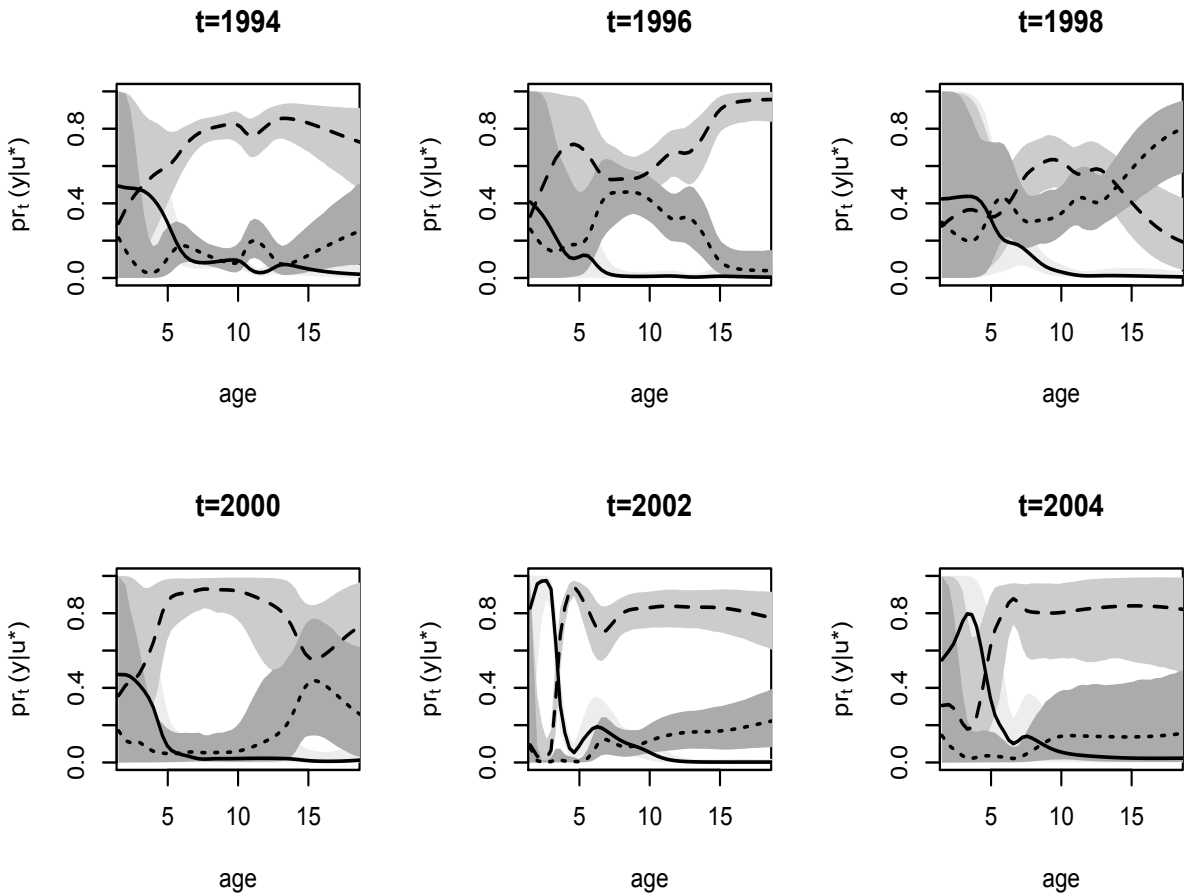
Posterior mean and 95% interval bands for the expected value of length over (continuous) age, across three years. Overlaid are the data (in blue) and the estimated von Bertalanffy growth curves (in red).

Rockfish data



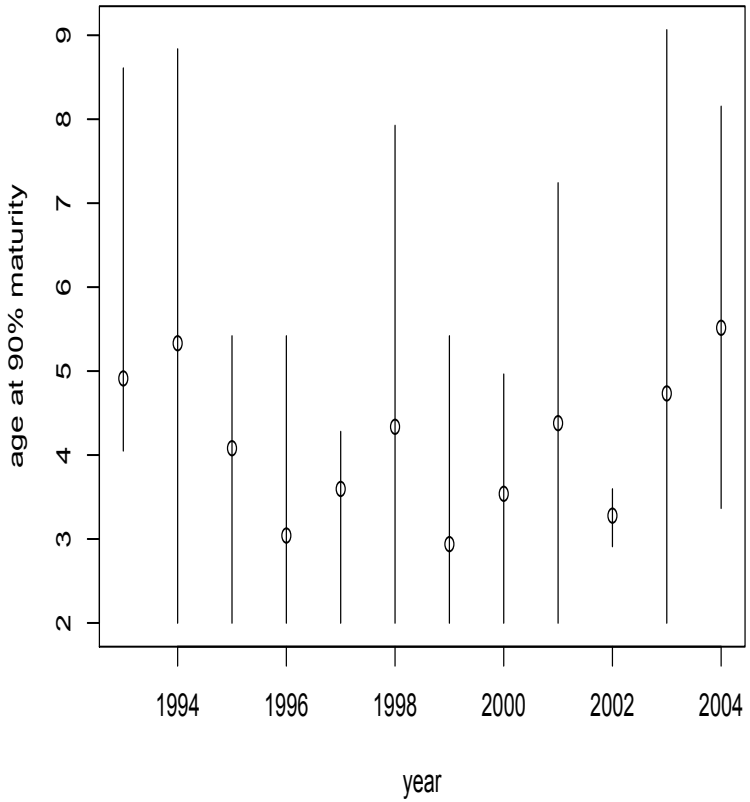
Posterior mean and 95% interval bands for the maturation probability curves associated with length: immature (solid); pre-spawning mature (dashed); post-spawning mature (dotted).

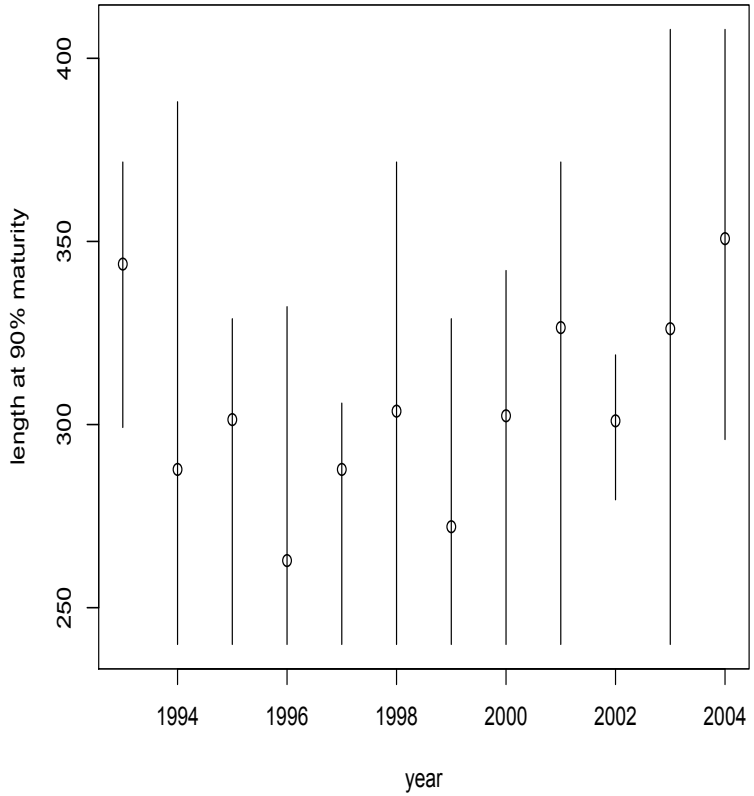
Rockfish data



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

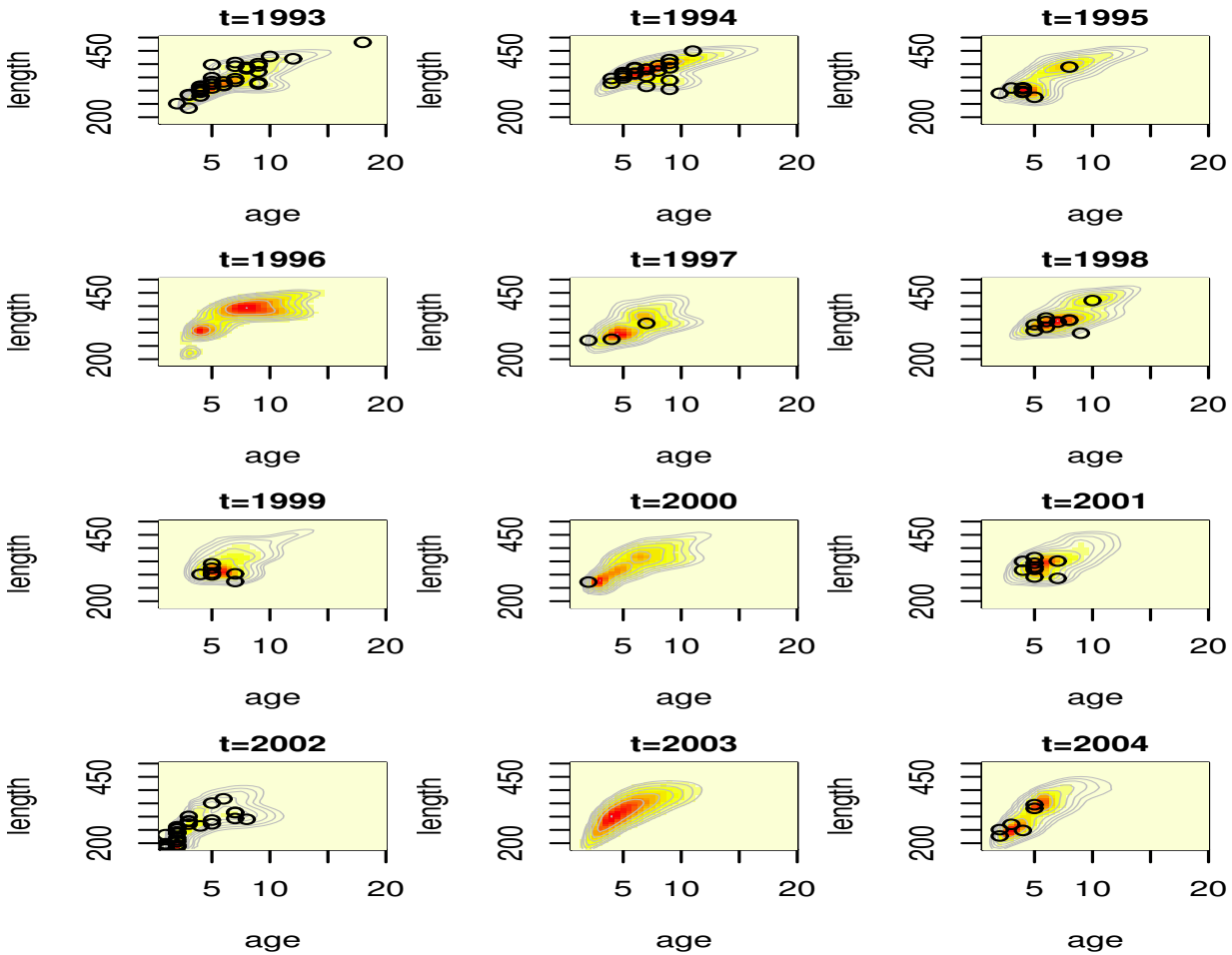
Rockfish data





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



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

Density regression in survival analysis

Density regression for survival responses

- Density regression approach in the context of survival analysis:
  - Non-standard shapes for the density (survival, hazard, mean residual life) function **and** non-linear regression relationships.
  - Survival analysis applications typically involve a small to moderate number of random covariates.
  - But, also covariates (e.g., binary control/treatment indicators) that need to be handled differently.

- DP mixture model for the joint response-covariate density:

$$f(t, \boldsymbol{x} \mid G) \quad = \quad \int k(t, \boldsymbol{x} \mid \boldsymbol{\theta}) \, \mathrm{d}G(\boldsymbol{\theta}) \approx \sum_{\ell=1}^{N} p_\ell \, k(t, \boldsymbol{x} \mid \boldsymbol{\theta}_\ell)$$

where $\boldsymbol{x}$ is the vector of (random) covariates, and the weighted mixture representation uses a truncation approximation, $G_N = \sum_{\ell=1}^{N} p_\ell \, \delta_{\boldsymbol{\theta}_\ell}$, to the DP stick-breaking construction for G.

Survival analysis functionals

Let T be an $\mathbb{R}^+$-valued random variable representing survival time.

- Survival function: $S(t) = \Pr(T > t)$.

- Hazard function: probability of failure in the next instant given survival up to time t, $h(t) = \lim_{\Delta t \to 0} \Pr[t < T \leq t + \Delta t \mid T > t]/(\Delta t)$
  - For continuous T, $h(t) = f(t)/S(t)$, where $f(t)$ is the density.

- Mean residual life (MRL) function: expected remaining survival time given survival up to time t. For continuous T,

$$m(t) = \mathrm{E}(T - t \mid T > t) = \frac{\int_t^\infty (u - t) f(u)\, \mathrm{d}u}{S(t)} = \frac{\int_t^\infty S(u)\, \mathrm{d}u}{S(t)}$$

  provided $\mu \equiv \mathrm{E}(T) = \int_0^\infty S(t)\, \mathrm{d}t < \infty$
  - The MRL function characterizes the survival distribution:

$$S(t) = \frac{m(0)}{m(t)} \exp\left[- \int_0^t \frac{1}{m(u)}\, \mathrm{d}u \right]$$

Functionals under the density regression model

- Mean regression:

$$\mathrm{E}(T \mid \boldsymbol{x}, G_N) = \sum_{\ell=1}^{N} q_\ell(\boldsymbol{x})\, \mathrm{E}(T \mid \boldsymbol{x}, \boldsymbol{\theta}_\ell)$$

where

$$q_\ell(\boldsymbol{x}) = p_\ell\, k(\boldsymbol{x} \mid \boldsymbol{\theta}_\ell) / \{\sum_{r=1}^{N} p_r\, k(\boldsymbol{x} \mid \boldsymbol{\theta}_r)\}$$

are covariate-dependent weights, and $\mathrm{E}(T \mid \boldsymbol{x}, \boldsymbol{\theta})$ is the conditional expectation under the mixture kernel distribution.

- Mean residual life regression:

$$m(t \mid \boldsymbol{x}, G_N) = \sum_{\ell=1}^{N} q_\ell(t, \boldsymbol{x})\, m(t \mid \boldsymbol{x}, \boldsymbol{\theta}_\ell)$$

where

$$q_\ell(t, \boldsymbol{x}) = p_\ell\, k(\boldsymbol{x} \mid \boldsymbol{\theta}_\ell) S(t \mid \boldsymbol{x}, \boldsymbol{\theta}_\ell) / \{\sum_{r=1}^{N} p_r\, k(\boldsymbol{x} \mid \boldsymbol{\theta}_r) S(t \mid \boldsymbol{x}, \boldsymbol{\theta}_r)\}$$

are covariate-dependent and time-dependent weights, and $m(t \mid \boldsymbol{x}, \boldsymbol{\theta})$ is the MRL function of the mixture kernel conditional response distribution.

Mixture kernel specification

- Condition for finite mean for the conditional survival distribution:
  - If $E_{G_0}[E(T \mid \boldsymbol{x}, \boldsymbol{\theta})] < \infty$, then $E(T \mid \boldsymbol{x}, G) < \infty$

- Gamma kernel component for the survival response variable:
  - Product kernel, $k(t, \boldsymbol{x}) = k(t)\, k(\boldsymbol{x})$ (with a gamma density for $k(t)$).
  - More general kernel: use appropriate marginal $k(\boldsymbol{x})$, and take

$$k(t \mid \boldsymbol{x}) = \Gamma(t \mid \exp(\theta), \exp(\boldsymbol{x}^T \boldsymbol{\beta}))$$

    such that $E(T \mid \boldsymbol{x}, \theta, \boldsymbol{\beta}) = \exp(\theta - \boldsymbol{x}^T \boldsymbol{\beta})$

DDP mixture model for treatment/control settings

Interest often lies in modeling survival times for treatment and control groups.

- Benefits in modeling dependence across groups.
- Let $s \in S$ represent the index of dependence, and consider $S = \{T, C\}$ where T and C are the treatment and control groups, respectively.
- DP mixture regression model:

$$f(t, \boldsymbol{x} \mid G_s) = \int k(t, \boldsymbol{x} \mid \boldsymbol{\theta}) \, \mathrm{d}G_s(\boldsymbol{\theta}), \quad s \in S$$

  where we seek to model the pair of dependent random mixing distributions (G_C, G_T).

- General DDP prior structure, $G_s = \sum_{l=1}^{\infty} \omega_{ls} \, \delta_{\boldsymbol{\theta}_{ls}}$, where marginally, $G_s \sim \mathrm{DP}(\alpha_s, G_{0s})$, for each $s \in S$.

DDP mixture model for treatment/control settings

- We might expect the two groups to be comprised of similar components, but possibly having varying prevalence, motivating modeling dependence only through the weights.

- We use mixing distribution, $G_s = \sum_{l=1}^{\infty} \omega_{ls}\, \delta_{\boldsymbol{\theta}_l}$, with a bivariate beta distribution defining the dependent stick-breaking weights (thus retaining the DP marginally).

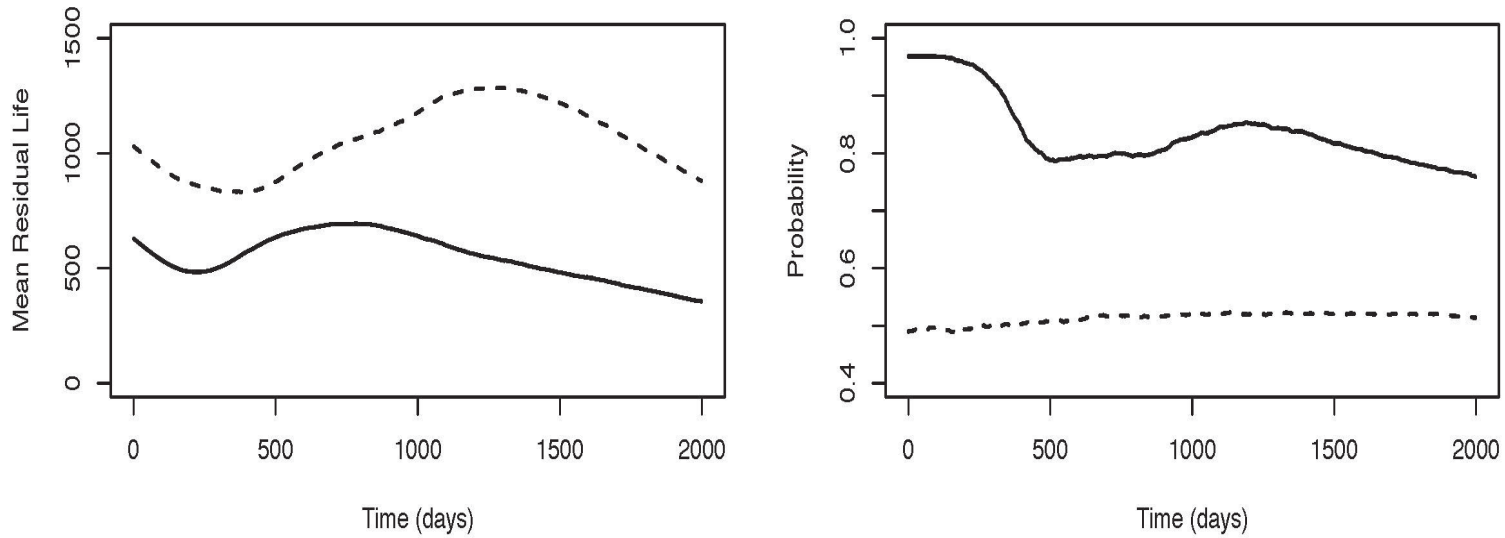
- With the truncated version of $G_s \approx \sum_{\ell=1}^{N} p_{\ell s}\, \delta_{\boldsymbol{\theta}_\ell}$, the model:

$$f(t, \boldsymbol{x} \mid G_s) = \int k(t, \boldsymbol{x} \mid \boldsymbol{\theta})\, \mathrm{d}G_s(\boldsymbol{\theta}) \approx \sum_{\ell=1}^{N} p_{\ell s}\, k(t, \boldsymbol{x} \mid \boldsymbol{\theta}_\ell), \qquad s \in \{T, C\}$$

- Practical benefit: modeling dependence only through the weights is not affected by the dimensionality of the mixture kernel.

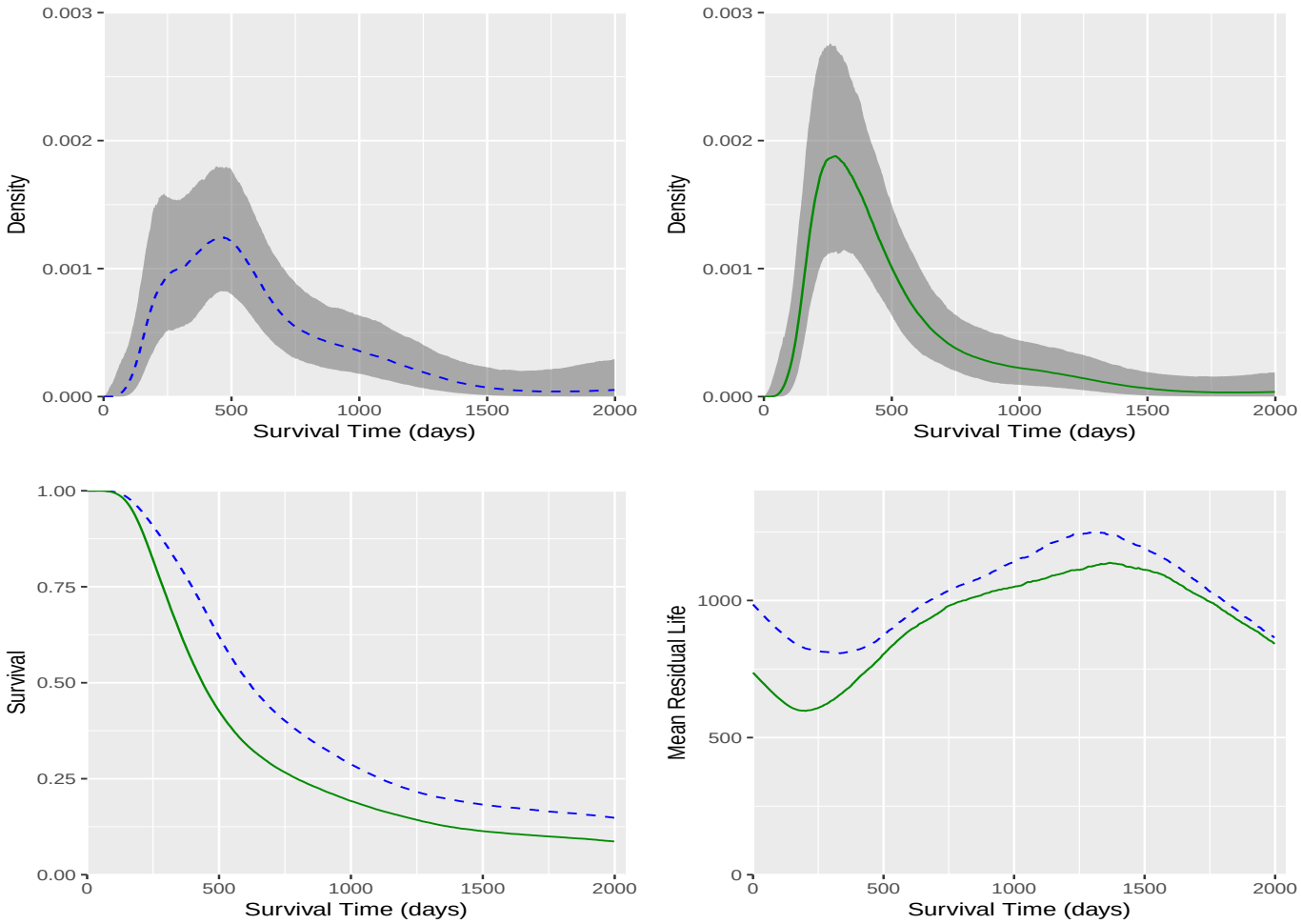
Small cell lung cancer data example

Study involving two treatments for small cell lung cancer (Ying et al., 1988): survival times (in days) for 121 patients (23 right censored) randomly assigned to one of two treatments $\rightarrow$ Arm A, under which 62 patients received cisplatin (P) followed by etoposide (E), and Arm B, where 59 patients received (E) followed by (P).



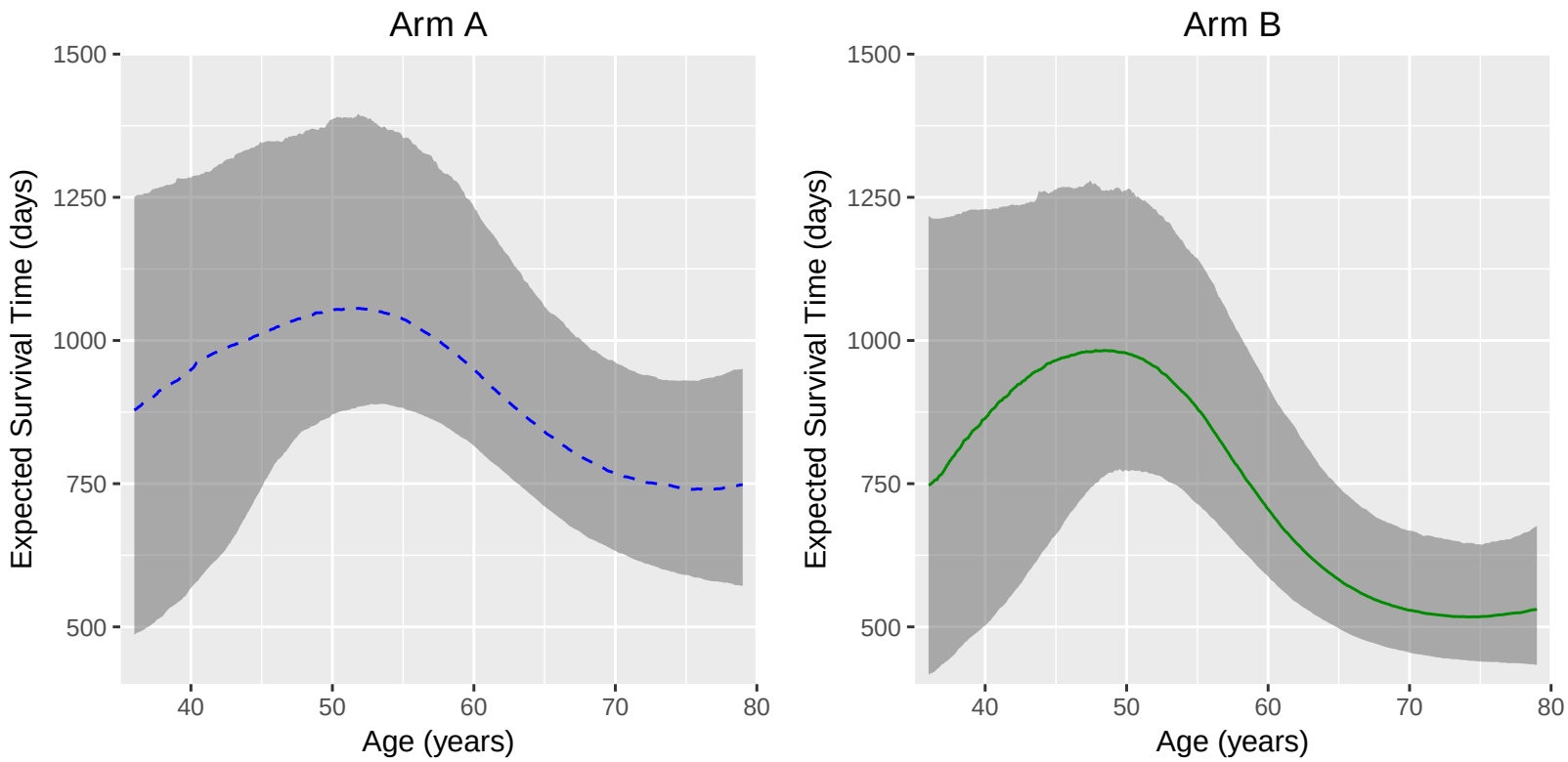
Left: posterior mean estimates for the Arm A (dashed line) and Arm B (solid line) MRL function. Right: $\Pr(m_A(t) > m_B(t))$ (dashed line) and $\Pr(m_A(t) > m_B(t) \mid \text{data})$ (solid line), as a function of time. (Results from a gamma DP mixture model applied separately to each group)

Small cell lung cancer data example



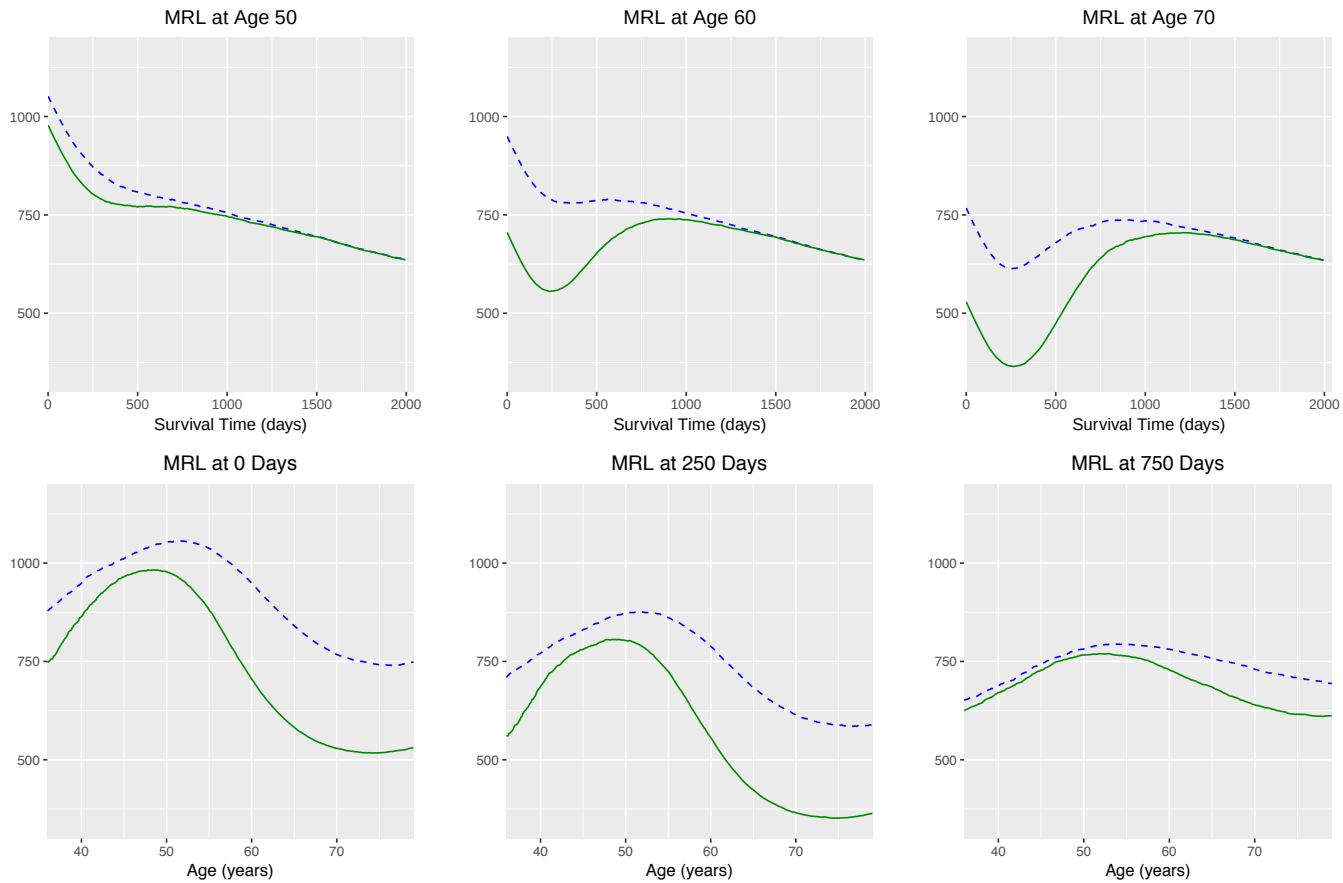
Point and 95% interval estimates of the density function for Arm A and Arm B. Point estimate of the survival function and the mean residual life function for Arm A (blue dashed) and Arm B (green solid). (Results from the gamma DDP mixture model)

Small cell lung cancer data example



Point and 80% interval estimates of the mean regression, $E(T \mid \text{age})$, for Arm A (blue dashed) and Arm B (green solid) across a grid of age values (in years). (Results from the gamma DDP density regression model, with the patient's age as the covariate)

Small cell lung cancer data example



Estimates of the MRL function of Arm A (blue dashed) and Arm B (green solid) for fixed ages (top panel) and for fixed times (bottom panels). (Results from the gamma DDP density regression model, with the patient's age as the covariate)

- Contact info:
  e-mail: thanos@soe.ucsc.edu
  web: https://users.soe.ucsc.edu/∼thanos

- Acknowledgment: funding from the NSF under awards SES 1950902 and DMS 2015428, and from the UCSC Committee on Research.

MANY THANKS!