A Short Course on Bayesian Nonparametrics

Lecture 1 - Introduction

Abel Rodriguez - UC, Santa cruz

Universidade Federal Do Rio de Janeiro
March, 2011

Bayesian paradigm

- Observations $y_1, \ldots, y_n$ are assumed to be generated by an unknown stochastic process F (Likelihood).
- Uncertainty about F *before observing* $y_1, \ldots, y_n$ is summarized in terms of a probability measure $P(F)$ (Prior)
- We can update $P(F)$ using Bayes theorem (Posterior)

$$P(F|y_1, \ldots, y_n) = \frac{F(y_1, \ldots, y_n)P(F)}{\int F(y_1, \ldots, y_n)dP(F)}$$

- Decisions are made by maximizing the expected utility $\hat{U}(d) = \int U(d, F)dP(F|y_1, \ldots, y_n)$.

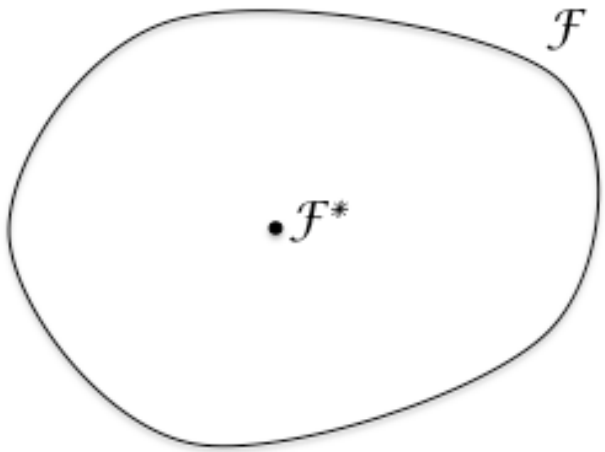
All models are wrong, but some are useful

Parametric vs. nonparametric Bayes

- **Parametric:** $y_i \in \mathcal{Y}$, $y_i \sim_{iid} F$, $F \in \mathcal{F}^*$,

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

  Prior on $\mathcal{F}^* \Rightarrow (\mu, \tau^2)$.
- $\mathcal{F}^*$ has measure 0 on $\mathcal{F}$ = the space of all measures on $\mathcal{Y}$!!!!
- **Nonparametric:** Priors on subsets of $\mathcal{F}$ that do not have measure 0 involve infinite-dimensional spaces $\Rightarrow$ Stochastic process.



Classical vs. Bayes Nonparametric

- Classical nonparametrics (ranks/order statistics)
  - Estimation: Empirical distribution / KDE.
  - Testing: Sign rank test.
- The term "Bayesian nonparametric models" is an oxymoron: we should really say "Bayesian models with an infinite number of parameters".
- Nonparametric or semiparametric? Proportional hazards models?

What I expect you to already know

- Some Bayesian parametric inference, including hierarchical modeling (for example, at the level of Gelman et al, 2003).
- Have some experience with simulation-based inference using MCMC (at the level of Gamerman & Lopes, 2006).
- A bit of measure theory (not a lot, mostly the basic concepts like σ-algebras).

Two traditional problems in nonparametric statistics

Regression

$$y_i = f(x_i) + \epsilon_i \quad \epsilon_i \text{ iid}, \mathrm{E}(\epsilon_i) = 0$$

with

$$f \in C^q \quad x_i \in \mathcal{S} \subset \mathbb{R}^p$$

$$f \sim Q_1$$

$$\Downarrow$$

Gaussian processes

Density estimation

$$y_i \sim F$$

with

$$y_i \in \mathcal{Y} \quad F \in \mathcal{F}$$

$$F \sim Q_2$$

$$\Downarrow$$

Dirichlet processes

What makes a non-parametric model "good"?

Raiffa and Schlaifer, 1961; 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 a large enough support.
- The hyperparameters in the model should be easily interpretable.

Semiparametric Regression

$$y_i = f(x_i) + \epsilon_i \quad \epsilon_i \sim \mathrm{N}(0, \tau^2)$$

- A prior for f involves specifying a distribution for every possible value of $f(x)$, with $x \in \mathcal{S} \subset \mathbb{R}^p \Rightarrow$ A stochastic process *on an uncountable index space.*

- Stochastic processes on uncountable spaces through joint distributions on countable collections $\Rightarrow$ Kolmogorov consistency conditions:
  - Symmetry.
  - Marginalization.

Gaussian Processes

Definition

$Z_{\mathcal{S}} = \{Z(s) : s \in \mathcal{S}\} \sim \text{GP}(\mu(s), \sigma(s, s'))$ iff for any n and $s_1, \ldots, s_n \in \mathcal{S}$

$$\begin{pmatrix} Z(s_1) \\ \vdots \\ Z(s_n) \end{pmatrix} \sim \text{N}\left(\begin{pmatrix} \mu(s_1) \\ \vdots \\ \mu(s_n) \end{pmatrix}, \begin{pmatrix} \sigma(s_1, s_1) & \cdots & \sigma(s_1, s_n) \\ \vdots & \ddots & \vdots \\ \sigma(s_n, s_1) & \cdots & \sigma(s_n, s_n) \end{pmatrix} \right)$$

- Widely used in spatial statistics!!!
- Brownian motion is a special case.

Gaussian processes

We observe $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ and assume

$$y_i | f, x_i \sim_{iid} \mathrm{N}(f(x_i), \tau^2) \quad \{f(x) : x \in \mathcal{S}\} \sim \mathrm{GP}(\mu(x), \sigma(x, x'))$$

or

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \mid \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix}, \tau^2 \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix} \right)$$

$$\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu(s_1) \\ \vdots \\ \mu(s_n) \end{pmatrix}, \begin{pmatrix} \sigma(s_1, s_1) & \cdots & \sigma(s_1, s_n) \\ \vdots & \ddots & \vdots \\ \sigma(s_n, s_1) & \cdots & \sigma(s_n, s_n) \end{pmatrix} \right)$$

or

$$y | f \sim \mathrm{N}(f, \tau^2 I) \qquad f \sim \mathrm{N}(\mu, \Sigma)$$

Gaussian processes

- Interpretability of parameters μ represents the prior "expected form" for f, and σ controls how close the realizations are to μ.
- Tractability: Conjugate analysis. Assuming you know μ and σ, the posterior for $f = (f(x_1), \ldots, f(x_n))'$ (the value of the function at the observed covariates) is

$$f|y \sim \mathrm{N}\left(\left\{\frac{1}{\tau^2} I + \Sigma^{-1}\right\}^{-1} \left\{\frac{1}{\tau^2} y + \Sigma^{-1}\mu\right\}, \left\{\frac{1}{\tau^2} I + \Sigma^{-1}\right\}^{-1}\right)$$

Gaussian processes

- Posterior at new locations, $x^*_{n+1}, \ldots, x^*_m$ is obtaining by exploiting properties of the multivariate normal

$$f_2|f_1 \sim \mathrm{N}(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}\{f_1 - \mu_1\}, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$$

- If $\tau^2 > 0$ this is a smoother, if $\tau^2 = 0$, this is an interpolator.
- In practice, τ^2, μ and Σ are unknown and are given hyperpriors $\Rightarrow$ We need MCMC for computation.

Gaussian processes vs. basis expansions

- Richness: There is a close connection between the "stochastic process" approach to nonlinear regression and basis rep.
- Kaurhunen-Loève representation: $f \sim \text{GP}(0, \sigma(s, s'))$ with $\sigma(s, s')$ jointly continuous and positive definite. Then

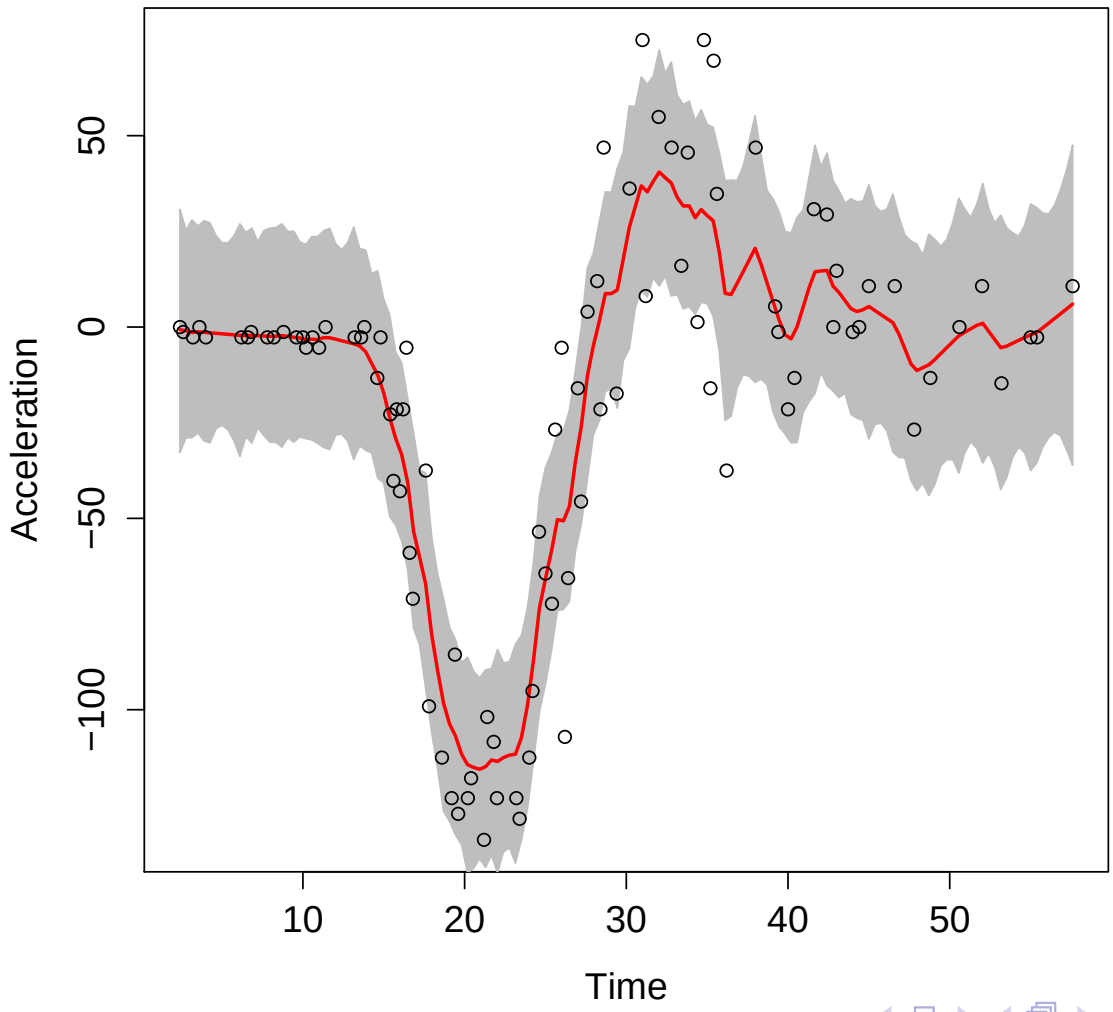
$$f(x) = \sum_{k=1}^{\infty} \beta_k \phi_k(x).$$

where $\beta_k \sim \text{N}(0, 1)$ independently and

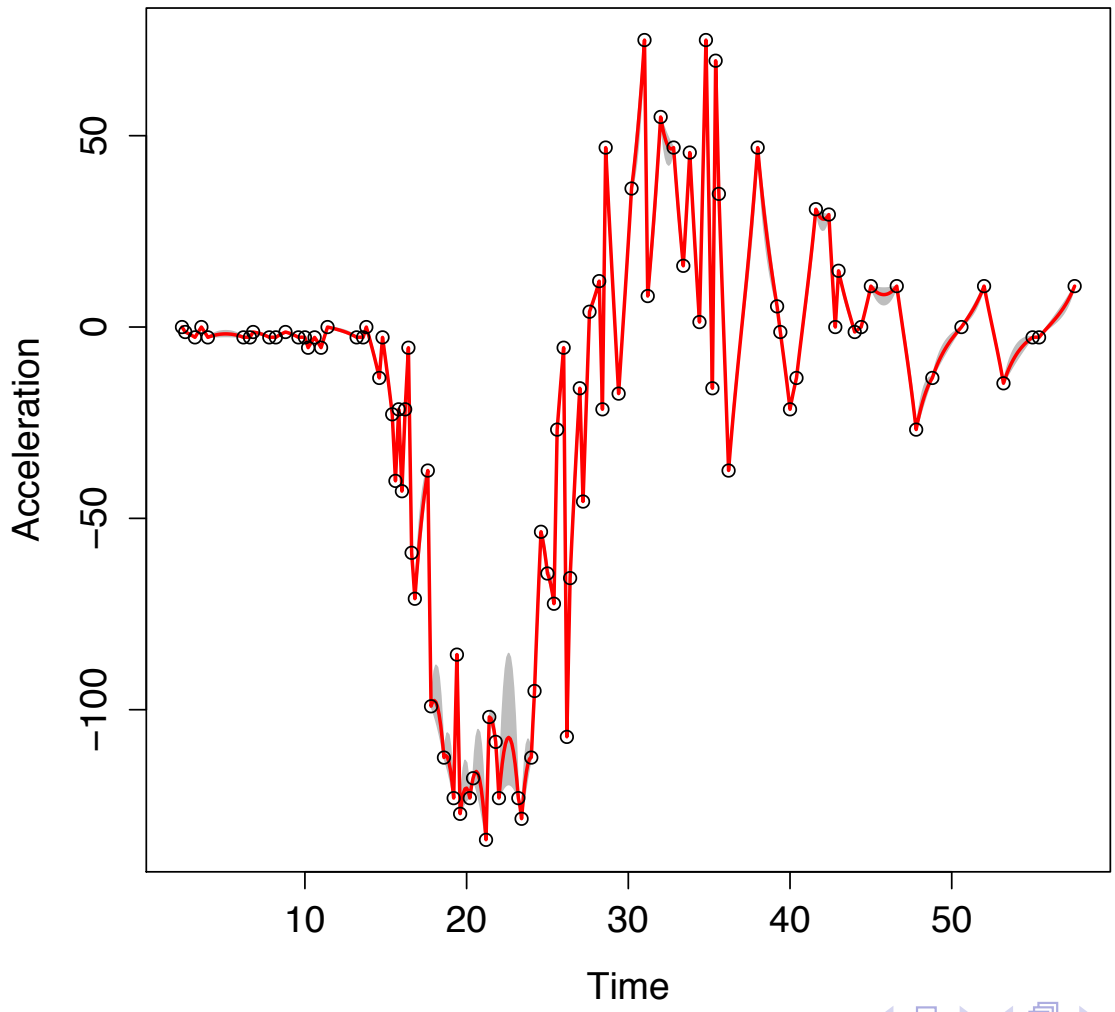
$$\lambda_k \phi_k(s) = \int \sigma(s, s') \phi_k(s') ds' \quad \int \phi_k(s) \phi_l(s) ds = \begin{cases} 1 & k = l \\ 0 & k \neq l \end{cases}$$

- Provides a link with splines, wavelets, Fourier, RKHS, etc...

The motorcycle data (Silverman, 1985) - Smoothing



The motorcycle data (Silverman, 1985) - Interpolation



Density estimation

We observe $y_1, \ldots, y_n$ with $y_i | F \sim F$.

- Samples from a GP: $\mathcal{S} \to \mathbb{R}$, but probability distributions: $\mathcal{S} \to [0, 1]$. Would you use a normal distribution to model a variance?
- Instead of working with $(F(y_1), \ldots, F(y_n))$, consider the increments of the process,

$$(F(y_1), F(y_2) - F(y_1), \ldots, 1 - F(y_n)) \in \text{Simplex}(\mathbb{R}^{n+1})$$

- The simplest prior on the simplex is the Dirichlet distribution.

$$(F(y_1), F(y_2) - F(y_1), \ldots, 1 - F(y_n)) \sim \text{Dir}(\alpha H(y_1), \alpha\{H(y_2) - H(y_1)\}, \ldots, \alpha\{1 - H(y_n)\})$$

Some properties of the Dirichlet distribution

- Start with independent $Z_j \sim \text{Gam}(a_j, 1)$, $j = 1, ..., k$ (with $a_j > 0$)
- Define

$$Y_j = \frac{Z_j}{\sum_{l=1}^{k} Z_l}$$

- Then $(Y_1, ..., Y_k) \sim \text{Dir}(a_1, ..., a_k)$ (singular w.r.t. Lebesgue measure on R^k, since $\sum_{j=1}^{k} Y_j = 1$)
- $(Y_1, ..., Y_{k-1})$ has density

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

- For $k = 2$, $\text{Dirichlet}(a_1, a_2) \equiv \text{beta}(a_1, a_2)$.

Some properties of the Dirichlet distribution

- Moments:

$$\mathrm{E}(Y_j) = \frac{a_j}{\sum_{l=1}^k a_l} \qquad \mathrm{E}(Y_j^2) = \frac{a_j(a_j+1)}{\left(\sum_{l=1}^k a_l\right)\left(1+\sum_{l=1}^k a_l\right)}$$

  and, for $i \neq j$,

$$\mathrm{E}(Y_i Y_j) = \frac{a_i a_j}{\left(\sum_{l=1}^k a_l\right)\left(1+\sum_{l=1}^k a_l\right)}$$

  (note that this implies negative correlations).
- Marginals follow Beta distributions, $Y_j \sim \text{beta}(a_j, \sum_{i \neq j} a_i)$. More generally,

$$(Y_1, \ldots, Y_{s-1}, \sum_{l=s}^k Y_l) \sim \text{Dir}(a_1, \ldots, a_{s-1}, \sum_{l=s}^k a_l)$$

The Beta distribution as a prior on distributions for binary data

- Let $y_{i,j} \in \{0, 1\} = \mathcal{Y}$ (binary data).
- In this case,

$$\mathcal{F} = \left\{ F(y) : F(y) = \theta \mathbf{1}_{[0,\infty)}(y) + (1 - \theta)\mathbf{1}_{[1,\infty)}(y), \theta \in [0, 1] \right\}$$

- Hence, to place a prior on $\mathcal{F}$ we only need to place a prior on $\theta = [0, 1]$.
- A natural choice is the Beta distribution (but certainly not the only one!!!).

The Dirichlet distribution as a prior on distribution for categorical data

- Let $y_{i,j} \in \{1, \ldots, k\} = \mathcal{Y}$ (categorical data).
- In this case,

$$\mathcal{F} = \left\{ F(y) : F(y) = \sum_{j=1}^{k} \theta_j \mathbf{1}_{[j,\infty)}(y), \theta \in \text{Simplex}(\mathbb{R}^k) \right\}$$

- A natural choice for a prior on Simplex($\mathbb{R}^k$) is the Dirichlet distribution.
- Note that the Dirichlet distribution makes the prior "consistent", in the sense that the joint distribution on subsets of the space is also Dirichlet.

Ferguson (1973) definition of the Dirichlet process

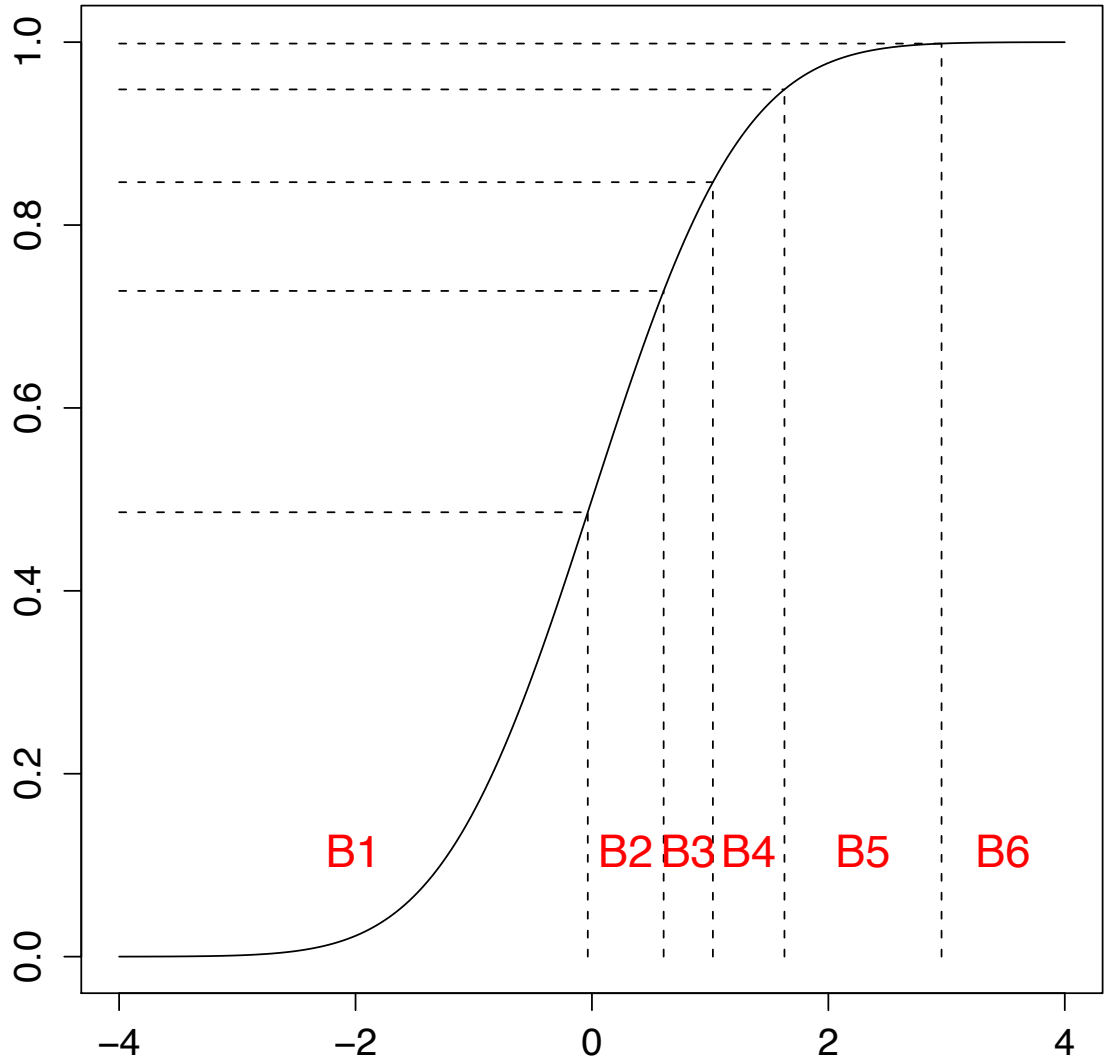
Definition

- $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}), F)$ is a probability space.
- $\mathcal{F}$ is the space of probability measures on $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$, so $F \in \mathcal{F}$.
- $F \sim \text{DP}(\alpha, H)$ iff for any n and any partition $B_1, \ldots, B_n$ of $\mathcal{Y}$

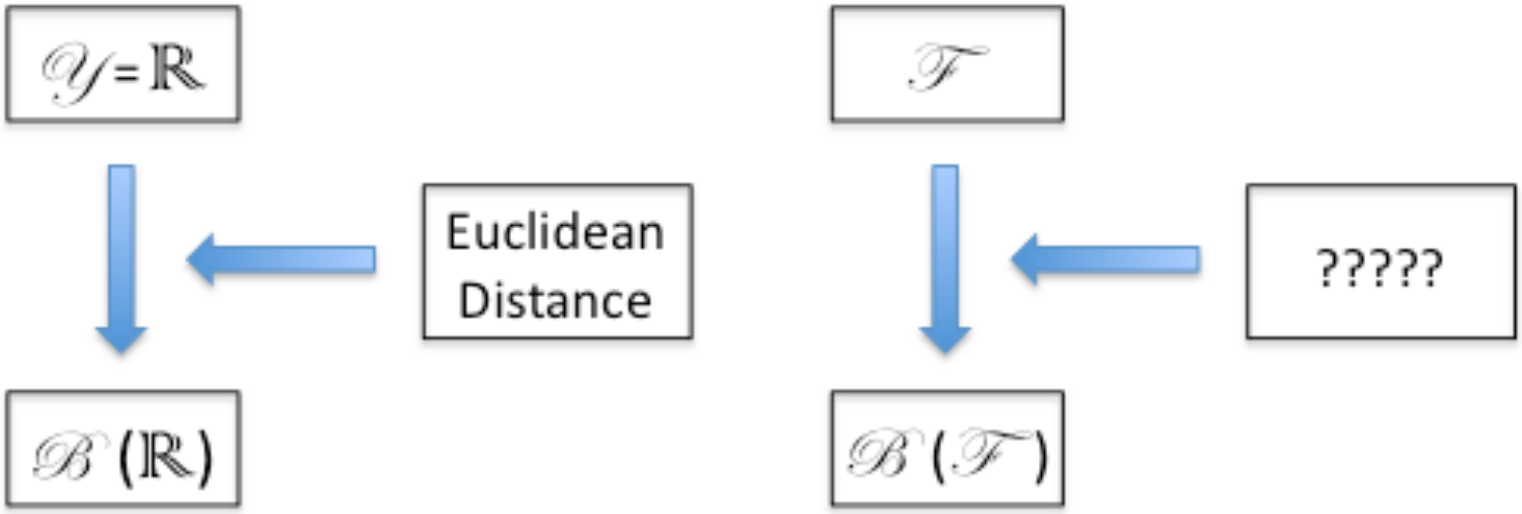
$$(F(B_1), F(B_2), \ldots, F(B_n)) \sim \text{Dir}(\alpha H(B_1), \alpha H(B_2), \ldots, \alpha H(B_n))$$

In many applications $\mathcal{Y} \subset \mathbb{R}^p$, so H and F are distribution functions $\Rightarrow$ I will often use the terms interchangeably.

The Dirichlet process



Some technical challenges



We need to think about $\mathcal{F}$ as a topological space.

What is an appropriate metric that we can use to define the Borel sets, and to explore issues like converge?

Density estimation using DP priors

- Interpretability: If $F \sim \text{DP}(\alpha, H)$, then for any measurable A,

$$\text{E}(F(A)) = H(A) \quad \text{Var}(F(A)) = H(A)\{1 - H(A)\}/(1 + \alpha)$$

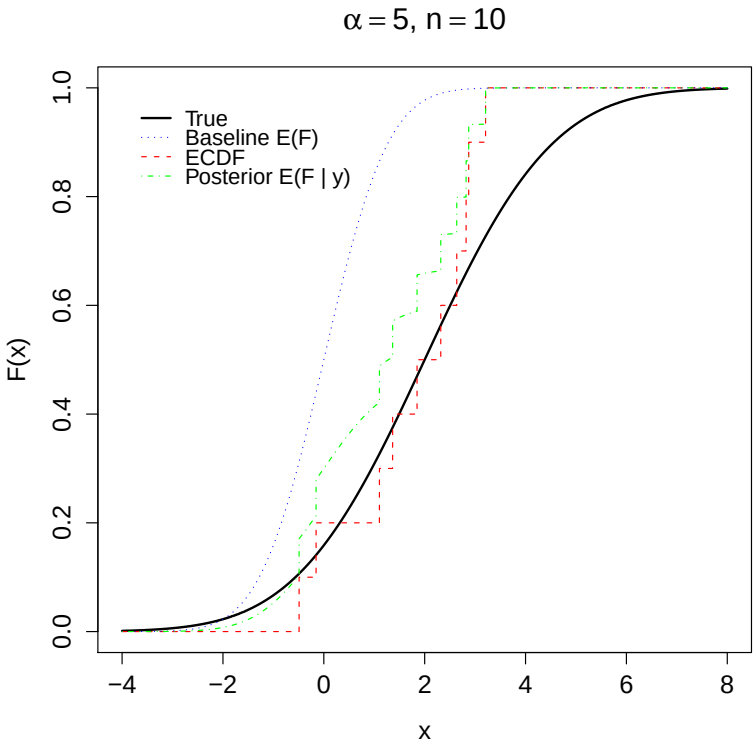
- Conjugate: If $y_i \sim_{iid} F$ for $i = 1, \ldots, n$ and $F \sim \text{DP}(\alpha, H)$

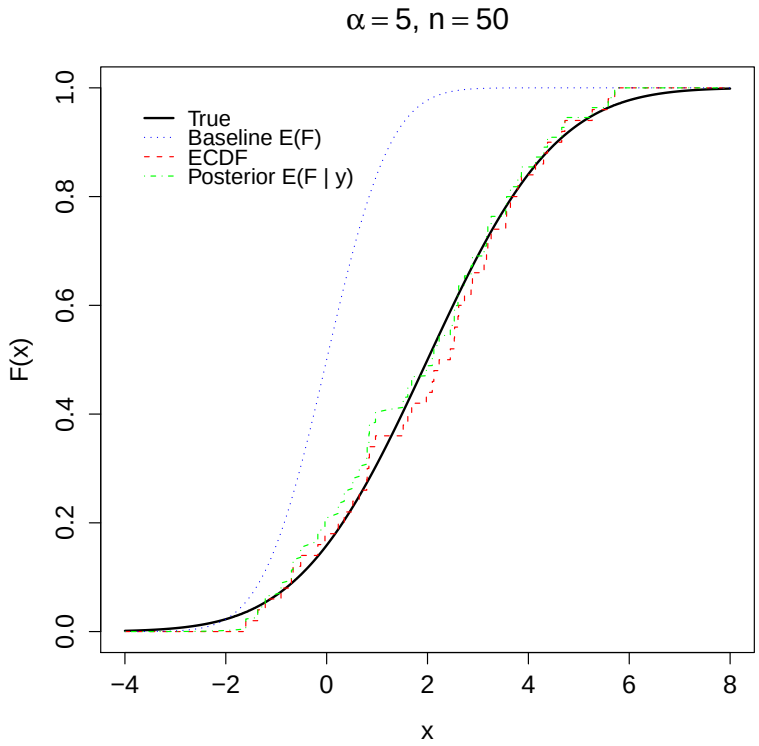
$$F|y \sim \text{DP}\left(n + \alpha, \frac{1}{n+\alpha}\sum_{i=1}^{n}\delta_{y_i} + \frac{\alpha}{n+\alpha}H\right) \Rightarrow$$

$$\text{E}(F(A)|y) = \frac{1}{n+\alpha}\sum_{i=1}^{n}\delta_{y_i}(A) + \frac{\alpha}{n+\alpha}H(A)$$

- Richness: Full support on a set that is dense on the space of measures that are absolutely continuous with respect to H.

Density estimation using DP priors





Constructive definition of the DP

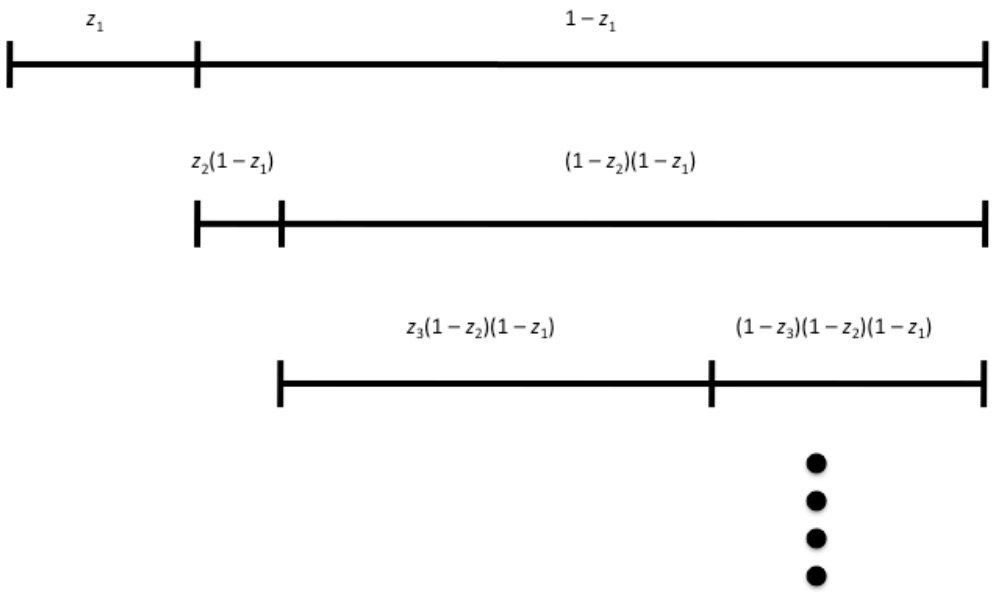
Definition

$F \sim \text{DP}(\alpha, H)$ iff

$$F(\cdot) = \sum_{k=1}^{\infty} \omega_k \delta_{\tilde{\vartheta}_l}(\cdot)$$

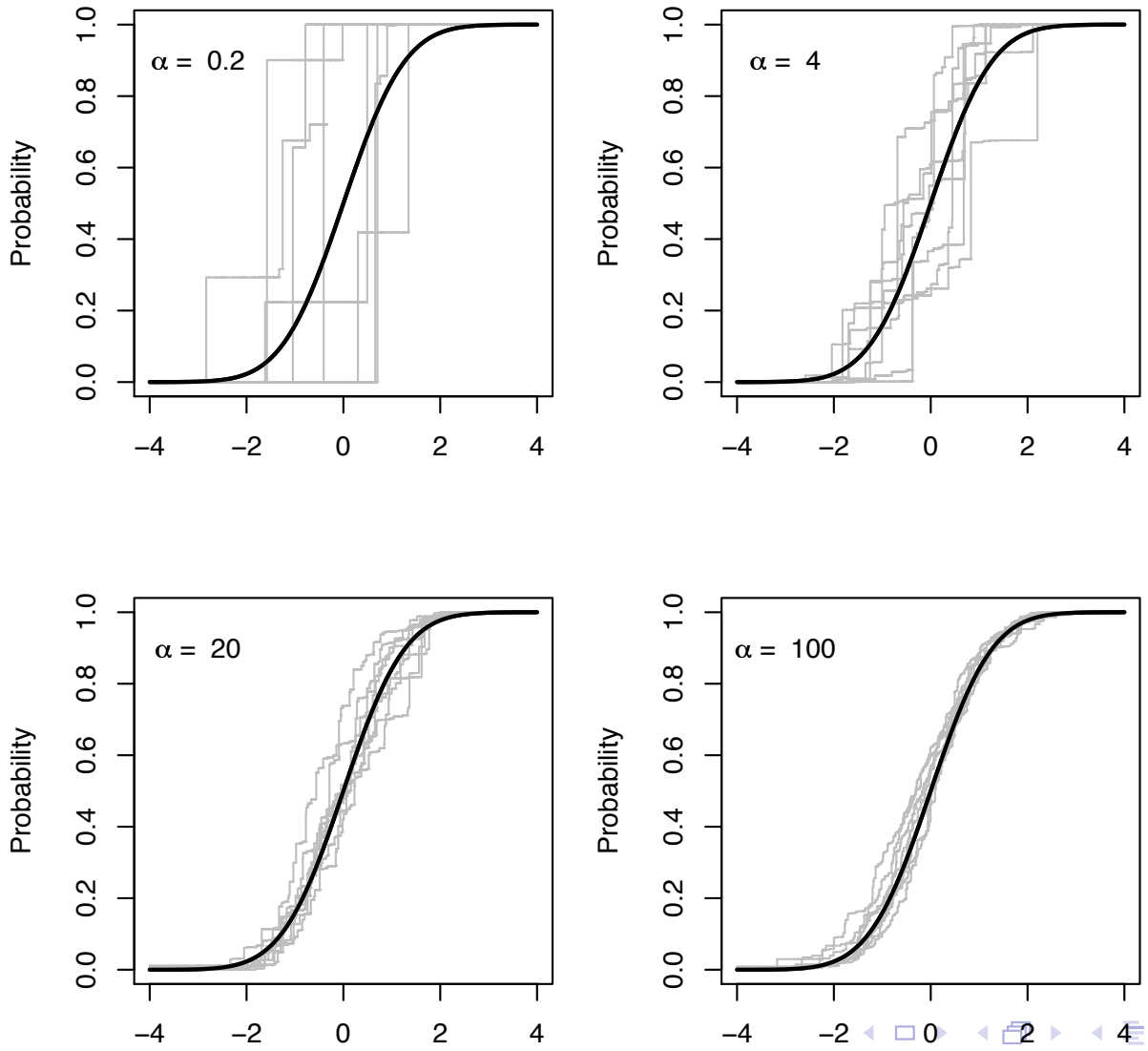
where $\tilde{\vartheta}_l \sim_{iid} H$, $\omega_k = z_k \prod_{l<k}\{1 - z_l\}$ and $z_l \sim_{iid} \text{beta}(1, \alpha)$.

Stick-breaking construction



- Note that $\sum_{k=1}^{\infty} \omega_k = 1$ almost surely (Ishwaran & James, 2001 have a general result).
- We say that $(\omega_1, \omega_2, \ldots) \sim \text{SB}(\alpha)$.
- Similar to continuation ratio models in survival analysis.

Samples from a Dirichlet process

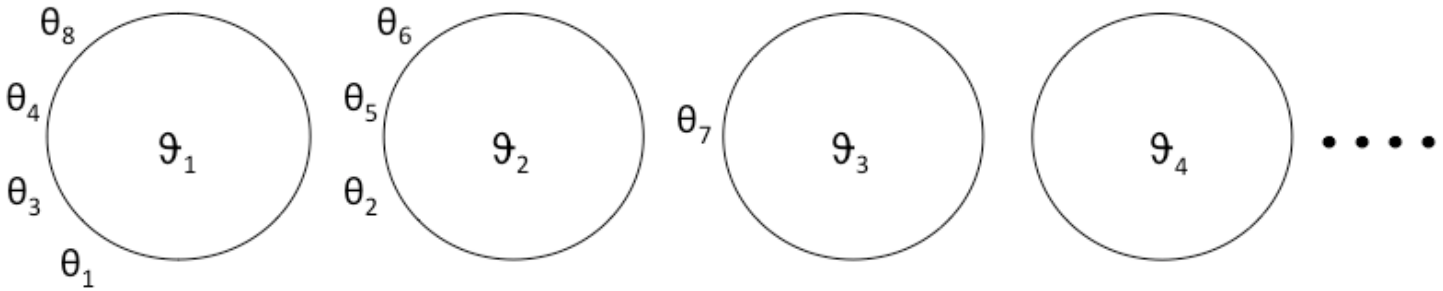


Pólya urn construction

If $\theta_i \sim_{iid} F$ and $F \sim \text{DP}(\alpha, H)$, then after integrating F,

$$P(\theta_i|\theta_{i-1}, \ldots, \theta_1) = \sum_{k=1}^{K^{i-1}} \frac{m_k^{i-1}}{n+\alpha}\delta_{\vartheta_l} + \frac{\alpha}{n+\alpha}H$$

We say $(\theta_1, \ldots, \theta_n) \sim \text{PU}(\alpha, H)$.



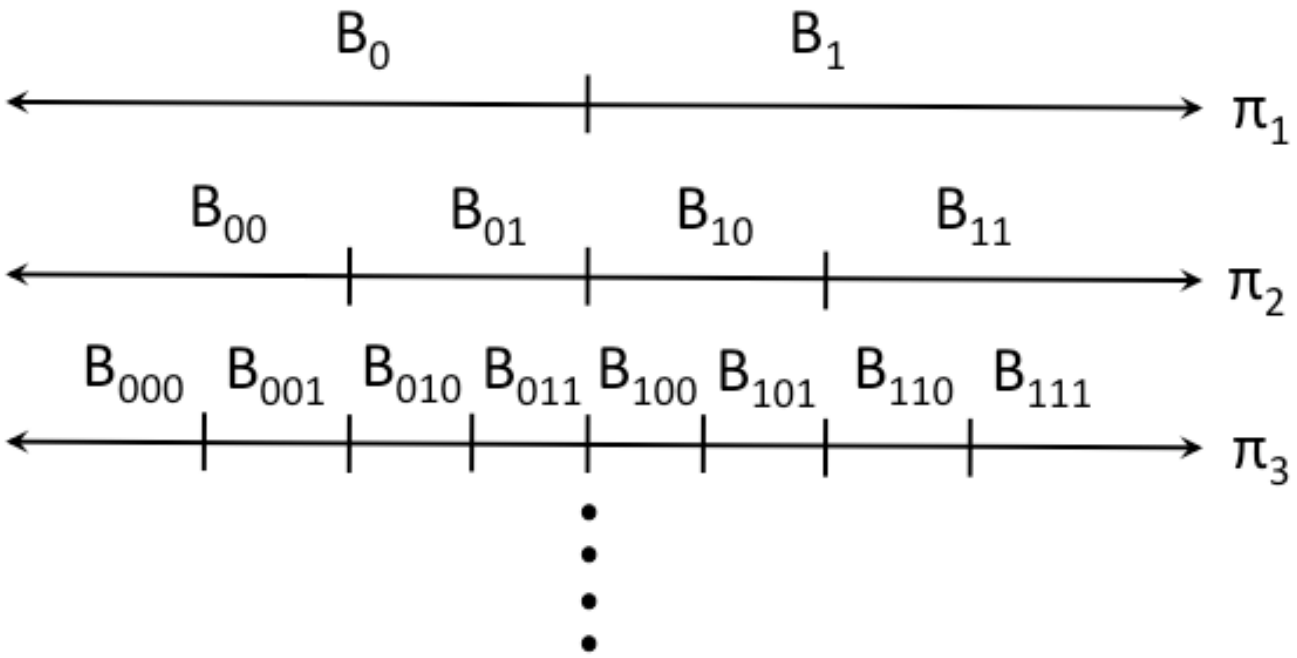
Correlation structure

- After integrating G, the θ_is are exchangeable (and hence have the same marginals and full conditional distributions), but they are not independent!!!
- Actually, for $i \neq j$ and $\theta_i \in \mathbb{R}$

$$\text{Cor}(\theta_i, \theta_j) = \frac{1}{1+\alpha} = \text{Pr}(\theta_i = \theta_j) = \text{Pr}(\theta_1 = \theta_2) = \sum_{k=1}^{\infty} \text{E}\{\omega_k^2\}$$

- Verify these identities, they will be useful later on.

Pólya Trees


B_0
B_1
π_1
B_{00}
B_{01}
B_{10}
B_{11}
π_2
B_{000}
B_{001}
B_{010}
B_{011}
B_{100}
B_{101}
B_{110}
B_{111}
π_3

Pólya Trees

Definition

A separating binary tree partition is a sequence of partitions $\Pi = \{\pi_t : t = 0, 1, \ldots\}$ such that $\bigcup_{k=0}^{\infty} \pi_k$ generates the measurable sets on $\boldsymbol{\Theta}$ and every $B \in \pi_{k+1}$ is obtained by splitting some $B^* \in \pi_k$ in two pieces.

Pólya Trees

- $D = \{0, 1\}$
- $D_0 = \emptyset$, $D_k = \underbrace{D \times D \times \cdots \times D}_{k}$ and $D^* = \bigcup_{k=0}^{\infty} D_k$.
- Π is a separating binary tree of partitions of $\boldsymbol{\Theta}$.
- $F \sim \text{PT}(\Pi, \mathcal{A})$, if there exists $\mathcal{A} = \{\alpha_\epsilon : \epsilon \in D^*\}$ and $\mathcal{Y} = \{Y_\epsilon : \epsilon \in D^*\}$ such that
  1. The random variables in $\mathcal{Y}$ are independent.
  2. For every $\epsilon \in D^*$, $Y_\epsilon \sim \text{beta}(\alpha_{\epsilon 0}, \alpha_{\epsilon 1})$.
  3. For every $k = 1, 2, \ldots$ and every $\epsilon_{1:k}$ we have

$$F(B_{\epsilon_{1:k}}) = \prod_{j=1;\epsilon_j=0}^{m} Y_{\epsilon_{1:j-1}} \prod_{j=1;\epsilon_j=1}^{m} (1 - Y_{\epsilon_{1:j-1}})$$

Dose response modeling

- Study potency of a stimulus by administering it at k dose levels to a number of subjects at each level
  - $\rightarrow x_i$: dose levels (with $x_1 < x_2 < ... < x_k$)
  - $\rightarrow n_i$: number of subjects at dose level i
  - $\rightarrow y_i$: number of $+$ responses at dose level i
- $F(x) = \Pr(\text{positive response at dose level } x) \Rightarrow$ dose-response curve.
- Standard assumption: the probability of a positive response increases with increasing dose level, i.e., F can be modeled as a cdf on $\mathcal{X} \subseteq R$.

Dose response modeling

Questions of interest:

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

Dose response modeling

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

Dose response modeling

- Assuming independent outcomes at different dose levels, the likelihood is given by

$$\prod_{i=1}^{k} \{F(x_i)\}^{y_i} \{1 - F(x_i)\}^{n_i - y_i}$$

- If the prior for F is a DP with precision parameter $\alpha > 0$ and base cdf H (the prior guess for the dose-response curve).
- The induced prior is

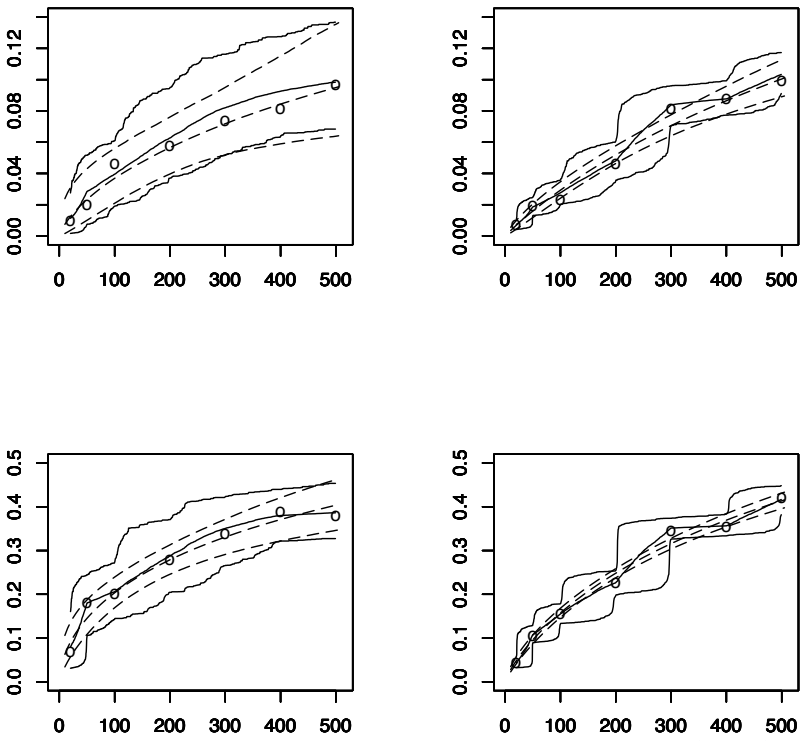
$$(F(x_1), F(x_2) - F(x_1), \ldots, 1 - F(x_k)) \sim$$
$$(\alpha H(x_1), \alpha\{H(x_2) - H(x_1)\}, ..., \alpha(1 - H(x_k)))$$

Dose response modeling

- The posterior for F is a *mixture of Dirichlet processes* (Antoniak, 1974) $\Rightarrow$ MCMC.
- The sampler augments the parameter space with $Z_{i,j} =$ number of subjects among the n_i receiving dosage x_i who have responded to it but not to dosage x_{i-1}.

$$Z_i = (Z_{i,1}, \ldots, Z_{i,k}) \sim \text{Mult}\{n_i, (F(x_1), F(x_2) - F(x_1), \ldots, 1 - F(x_n))\}$$

Dose response modeling



Example: Subjective elicitation of expert opinions

- Modeler $\neq$ Expert
- The modeler asks the expert questions like:

"What is the probability that we get a storm that leaves more than x inches of snow this winter? "

x	$\tilde{S}(x)$
2	0.70
3	0.60
4	0.50
5	0.30
7	0.05

- The goal is to elicit a distribution function $F(x) = 1 - S(x)$ that generated the observations $\tilde{F}(x_1), \ldots, \tilde{F}(x_n)$, where $\tilde{F} = 1 - \tilde{S}$

Example: Subjective elicitation of expert opinions

- Parametric: $F(x) \in \{G(x|\theta) : \theta \in \Theta\}$. Use the values of $\tilde{F}(x_1), \ldots, \tilde{F}(x_n)$ to create estimates for θ (medians, IQR, etc). How to reconcile discrepancies?
- Nonparametric (West, 1988): The *modeler* provides a prior guess H for the shape of the distribution, which is combined with the observation (the expert opinion). If $F \sim \text{DP}(\alpha, H)$,

$$(\tilde{F}(x_1), \tilde{F}(x_2) - \tilde{F}(x_1), \ldots, 1 - \tilde{F}(x_n)) \sim \text{Dir}(\alpha, H)$$

  (similar to an interpolating GP) Is the expert really sure?
- Gaussian process priors on the density associated with F (Oakley and O'Hagan, 2007). Can incorporate uncertainty, but prior has wrong support.

Example: Subjective elicitation of expert opinions

An alternative (the parametric case)

- Assume that $F(x) \in \{H_\theta(x) : \theta \in \Theta\}$. Hence the focus is on eliciting θ!!!
- The values of $\tilde{F}(x_1), \ldots, \tilde{F}(x_n)$ are noisy versions of the true $F(x_1), \ldots, F(x_n)$.
- The modeler provides H and a prior guess for θ.
- The model is:

$$\tilde{F}|\theta \sim \text{DP}(\alpha, H_\theta) \qquad\qquad \theta \sim p(\theta)$$

- Posterior sampling for θ using MCMC.
- The model is similar, but not the same as West, 1988!!!!

Example: Subjective elicitation of expert opinions

An alternative (the nonparametric case)

- Assume that $F(x) \in \mathcal{F}$=all distribution on $\mathcal{X}$
- As before, the values of $\tilde{F}(x_1), \ldots, \tilde{F}(x_n)$ are assumed to be noisy versions of the true $F(x_1), \ldots, F(x_n)$.
- As West, 1988, the modeler provides a guess H for F.
- The model is:

$$\tilde{F}|F \sim \text{DP}(\alpha, F) \qquad\qquad F \sim \text{DP}(\alpha, H)$$

- Akin to a smoothing GP.
- Identifiability in a single-expert/multiple-expert setting?

Homework:

Take the example in Section 5.2 of Moala and O'Hagan (2010) and apply the West (1988) and the parametric approach with uncertainty with

$$F(x) \in \{\text{LogNormal}(\phi, \Lambda), \phi \in \mathbb{R}^2, \Lambda \text{ Positive definite}\}$$