# Bayesian Nonparametrics: An Overview

Wesley Johnson

Department of Statistics, UC Irvine

SAMSI July 2010

# References


- Hanson, Branscum and Johnson (2005). Bayesian Nonparametric Modeling and Data Analysis: An Introduction. *in* **Handbook of Statistics.** Elsevier.
- Christensen, Johnson, Branscum and Hanson (2010). **Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians** (Chapter 15). CRC Press.
- Hanson and Jara (2009). Unpublished Lecture Notes
- Müller (2009). Unpublished Lecture Notes

# Why Bayesian Nonparametrics?

- Because parametric models are often overly restricted and/or lack robustness!
- So that we can find *biological bumps* that we might not otherwise find!
- So that we can see if parametric models might actually fit by embedding them in NP families!
- Because Bayesian NP modeling is feasible due to modern MCMC methods eg. *because we can*?

# Bayesian Parametric Models

- Given data $x = (x_1, ..., x_n)$ we model them with a joint pdf

$$Pr(X \in A \mid \theta) = \int_A f(x \mid \theta)\mu(dx) \quad \theta \in \Theta \subset R^k$$

- We treat the data as fixed and known and use the likelihood function to inform us about $\theta$

$$L(\theta) \propto f(x \mid \theta)$$

- We model our uncertainty about unknown $\theta$ through the use of a prior pdf, $p(\theta)$, which must be based on information that is independent of $x$ (failing in this results in Empirical Bayes methods)

# Parametric Inference

- Bayesian inference is facilitated through calculation of posterior pdf

$$p(\theta \mid x) = \frac{L(\theta)\, p(\theta)}{\int_{\Theta} L(\theta)\, p(\theta) d\theta}$$

- Due to intractability of integration, we use Markov chain Monte Carlo methods to sample from the joint posterior
  - Gibbs Sampling
  - Metropolis Sampling
  - Slice Sampling
  - Adaptive Rejection Sampling
  - Hybridizations of the above

# Parametric Inference

- We approximate integrals by

$$\int g(\theta)p(\theta \mid x)d\theta \doteq \sum_{i=1}^{MC} g(\theta^i)/MC$$

$$\theta^i \overset{iid}{\sim} p(\theta \mid x)$$

- So the posterior mean (vector) is numerically approximated as the arithmetic average of samples from the joint posterior.
- We obtain approximate 95% Probability Intervals for $\gamma \equiv g(\theta)$ by ordering $\{\gamma^i = g(\theta^i) : i = 1, ..., MC\}$ from smallest to largest and finding the 0.025 and 0.975 sample percentiles.
- The post med of $\gamma$ is more sensible than the post mean

# Wonderful Aspects

- Appropriateness of methods doesn't depend on having large sample sizes
- General ability to handle complex models without having to fall on mathematical swords
- Availability of statistical software is no longer an issue eg. WinBUGS, Open Bugs, JAGS, SAS, DP-Package etc.
- Inferences for complicated functions of $\theta$, eg.$\gamma = g(\theta)$, are available for the asking
- Direct probability interpretations

# The Sword We Do Have to Fall On

- The prior needs to be specified
- The more complex the model, the greater the potential difficulty in specifying a prior that will lead to a proper posterior (Hobert and Casella, JASA, 1996)
- Convergence of MCs can be challenging
- Some users of Bayesian statistics search for priors that will result in convergence of Markov chains
- With smaller sample sizes, the priors can matter a lot
- Even with large sample sizes, the priors can matter
- Sensitivity analysis and appropriate selection of prior is important

# Types of BNP Modeling

- A standard semi-parametric regression model is the simple linear model, only without the assumption of a parametric family for the errors

$$y_i = x_i\beta + \varepsilon_i \qquad \varepsilon_i \overset{iid}{\sim} P \quad P \in \mathbf{P}$$

where $\mathbf{P}$ is a large family of (preferably median 0) distributions, possibly including the Normal. The problem becomes Bayesian when we place a prior on $\mathbf{P}$

- Standard non-parametric models might simply assert

$$\begin{array}{ll} (i) & x_i \mid P \overset{iid}{\sim} P \qquad P \in \mathbf{P} \\ (ii) & x_i \mid P_x \overset{iid}{\sim} P_x \qquad y_j \mid P_y \overset{iid}{\sim} P_y \qquad P_x \perp P_y \\ & x_i \perp y_j \qquad (P_x, P_y) \in \mathbf{P_x} \times \mathbf{P_y} \end{array}$$

# BNP Modeling

- A NP regression model might specify

$$y_i \mid x_i, P_{x_i} \overset{ind}{\sim} P_{x_i} \qquad \{P_{x_i} : i = 1, ..., n\} \in \mathbf{P_X}$$

  where the prior on $\mathbf{P_X}$ allows the $P_{x_i}$ s to be correlated.
- So this model requires a distribution on multiple large families of distributions.
- This generality in principle allows one to estimate the regression functions eg. $E(y \mid x)$, $\text{Var}(y \mid x)$, as well as the density functions $f(y \mid x)$.

# Mean Regression Modeling

- An entirely separate area involves the model

$$y_i = m(x_i) + \varepsilon_i \qquad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

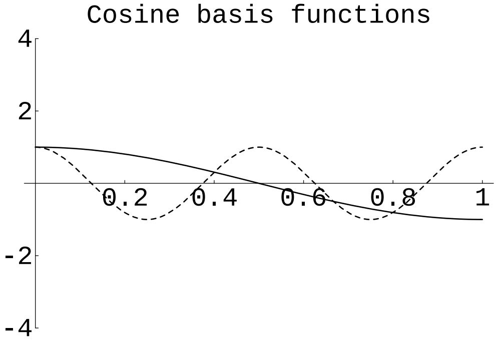
  but where $m(\cdot)$ is arbitrary.
- Usually, $m(\cdot)$ is modeled as

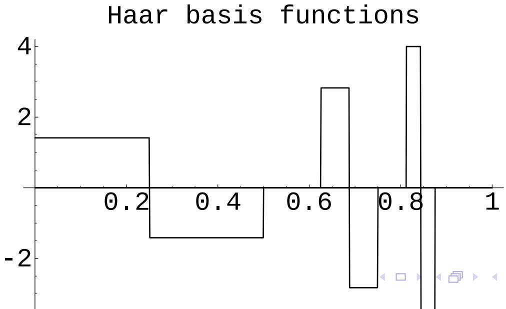
$$m(x) = \beta_0 + \sum_{k=1}^{\infty} \beta_k \phi_k(x)$$

  where the $\phi_k(\cdot)$ s form a basis for the space spanned by functions like $m(\cdot)$.
- Typical basis functions are Wavelets, B-splines, splines etc. They are of course truncated and much effort is given to the topic of “thresholding” in the literature.
- A typical Bayesian model places priors on the regression coefficients that allows for point masses at 0, which handles the thresholding

# Mean Regression Modeling

- A typical Bayesian model places priors on the regression coefficients that allows for point masses at 0, which handles the thresholding
- Ethanol Data: Response $y$ is the amount of nitric oxide and dioxide from a single engine in micrograms per joule, and the predictor, $x$, is a measure of he air to fuel ratio.
- We give estimates of the mean regression function using Cosine, Haar and B-spline basis functions truncated at $K$.


Cosine basis functions
4
2
0.2
0.4
0.6
0.8
1
-2
-4


Haar basis functions
4
2
0.2
0.4
0.6
0.8
1
-2

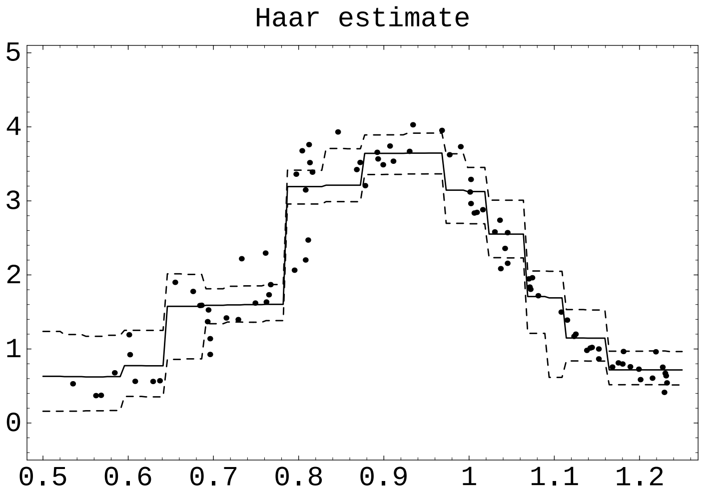
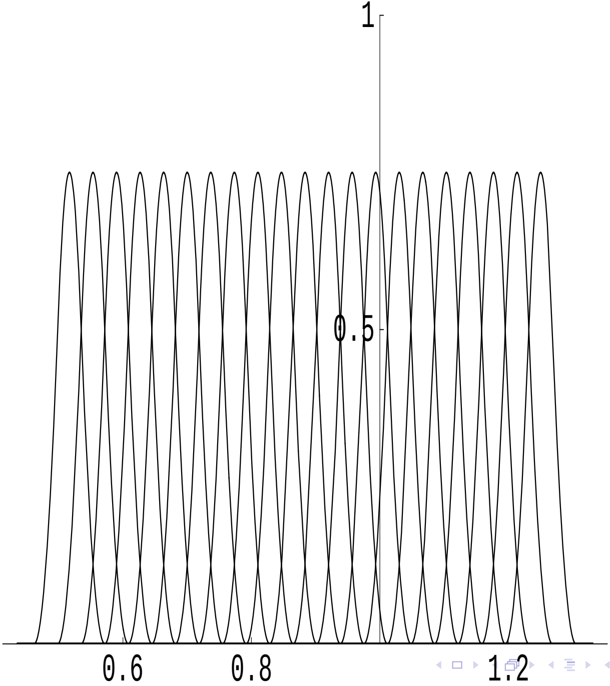



Figure 15.10 *Ethanol Data: Estimates of regression mean functions using Harr Wavelets.*


1
0.5
0.6
0.8
1.2

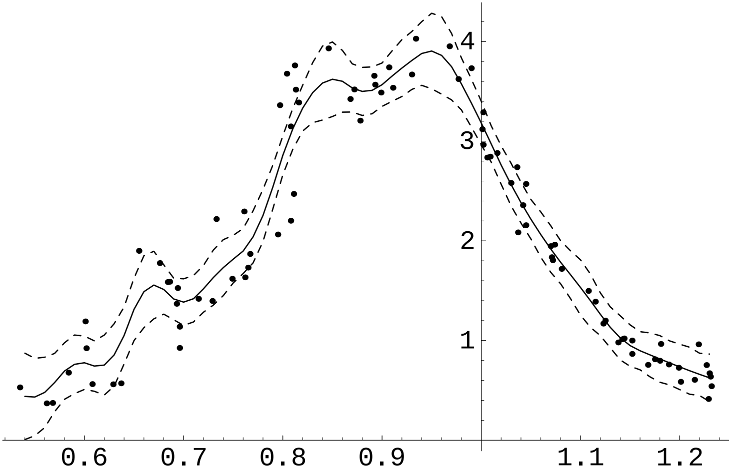



Figure 15.13 *Estimated trend using quadratic B-splines with* $K = 21$ *knots.*

# Popular Non-Parametric Priors

- Dirichlet Process (Ferguson, 1973)
- Dirichlet Process Mixtures (Lo, AOS 1994; Escobar, JASA 2004; Escobar and West, JASA 2005)
- Mixtures of Dirichlet Processes (Antoniak, AOS 1974, Berry and Christensen, AOS 1979; Hanson and Johnson, JCGS 2002)
- Mixtures of Polya Trees (Lavine, AOS 1992, 1994; Berger and Guglielmi, JASA 2001; Hanson and Johnson, JASA 2002; Hanson, JASA 2006)

# Dirichlet Process (DP)

- Sethuraman (1994)

$$P \mid F_0, c \sim DP(c, G_0)$$

- $\Leftrightarrow$

$$P = \sum_{h=1}^{\infty} p_h G_{\theta_h}(\cdot)$$

$$p_h = u_h \prod_{j=1}^{h-1} (1 - u_j) \quad u_h \overset{iid}{\sim} \text{Beta}(1, c)$$

$$\theta_h \overset{iid}{\sim} G_0$$

# Dirichlet Process as a Model for Data

- *Bad*, since discrete with probability one
- In the iid data case, the posterior mean behaves like the Empirical CDF (Susarla and VanRyzin, circa 1977) or like Kaplan-Meier in censored data case
- $E(G) = G_0 \Rightarrow$ prior is centered on the specific prior guess $G_0$. Not so good.
- Bruce Hill was often quoted in the 1980's that "there should have been only one paper written on the DP" (eg. Ferguson 1973)
- If $X \sim G \Rightarrow$ then $Pr(X \in A) = G_0(A)$ eg. marginal for $X$ is $G_0$
- Conjugacy: $G \mid X = x \sim DP(c + 1, \frac{c}{c+1} G_0(\cdot) + \frac{1}{c+1}\delta_x(\cdot))$

# Mixture of Dirichlet Processes (MDP)

- Let $G_0 = G_\theta$, a parametric model, and specify $p(\theta)$
- Then write
$$X \sim \int DP(c, G_\theta) p(\theta) d\theta$$
eg. Mixture of DPs
- When $c$ is large, the model tends to the parametric model $G_\theta$ with a standard prior on $\theta$ eg. $p(\theta)$
- When $c$ is small, we have a large family of possible distributions that includes the parametric family.
- Since $\mathrm{E}[G(\cdot) \mid \theta] = G_\theta(\cdot)$ for all $\theta$, we have centered the NP prior on the specified parametric family

# Dirichlet Process Mixture

- We say $X$ is drawn from a DPM if:

$$\begin{aligned} X \mid \theta &\sim G_\theta \\ \theta \mid G &\sim G \\ G \mid G_0, c &\sim DP(c, G_0) \end{aligned}$$

$$f(x \mid G) = \int f(x \mid \theta) dG(\theta) = \sum_{h=1}^{\infty} p_h f(x \mid \theta_h)$$

$$\theta_h \overset{\text{iid}}{\sim} G_0$$

- Replace $G_0$ with $G_\gamma$ and incorporate prior $p(\gamma)$ eg. Mixture of DPMs

# Dirichlet Process Mixture

- $E[F(x) \mid G_0] = \int F(x \mid \theta) dG_0(\theta)$
- For large $c$,

$$f(x \mid G) \doteq \int f(x \mid \theta) dG_0(\theta)$$

- So $G_0$ behaves like a prior for $\theta$ in the large $c$ (parametric) case
- But it's not the same as centering the NP model on a parametric family
- Expected number of terms in the mixture is approx $c \ell n(\frac{c+n}{c})$; can be small eg. 5 when $c = 1, n = 150$

# Dirichlet Process Mixture

- The DPM is by far the most popular NP model for data
- The Bayesian part involves choice of $G_0$ (or $G_\gamma$) and $c$
- Prior is often placed on $c$ (Escobar and West, 1995)
- Standard $G_0$ in the case of normal $G_\theta$ family is the usual conjugate prior eg Normal-Gamma
- Often, rather than selecting parameter values for Normal-Gamma, further priors placed on these
- Subjective priors not used in my limited experience
- “Non-informative” priors and/or resort to Empirical Bayes

# Marginalized DPM

- Early and perhaps most inferences through marginalization eg

$$f(x_i) = \int \int f(x_i \mid \theta_i) dG(\theta_i) dP(G)$$

- The $x_i$ s are (jointly) exchangeable
- Gibbs sampling entails sampling $\theta_i \mid \theta_{(i)}, x$ using the (updated) Polya Urn scheme

$$\begin{aligned} p_{\theta_i}(\theta \mid \theta_{(i)}, x) &= \frac{c f(x_i \mid \theta_i) dG_0(\theta) + \sum_{j \neq i} f(x_i \mid \theta) \delta_{\theta_j}(\theta)}{c \int f(x_i \mid \theta) dG_0(\theta) + \sum_{j \neq i} f(x_i \mid \theta_j)} \\ &\equiv q_0 p_{\theta_i}(\theta \mid x_i) + \sum_{j \neq i} q_j \delta_{\theta_j}(\theta) \end{aligned}$$

- From this, the seed is planted for the development of random partition models

# Marginalized DPM: Predictive Density

- Let $\theta = \{\theta_i : i = 1, ..., n\}$. Then

$$
\begin{aligned}
f(x_{n+1} \mid x) &= \int f(x_{n+1} \mid \theta, \theta_{n+1}, x) p(\theta_{n+1}, \theta \mid x) d\theta\theta_{n+1} \\
&= \int f(x_{n+1} \mid \theta_{n+1}) \int [p(\theta_{n+1} | \theta) p(\theta | x) d\theta] d\theta_{n+1}
\end{aligned}
$$

- The above can be numerically approximated by taking the Gibbs Sample of $\theta^j : j = 1, ..., MC$; then sample $\theta^j_{n+1}$ from the Polya Urn scheme

$$
\theta_{n+1} \mid \theta \sim \frac{c}{c+n} G_0(\cdot) + \frac{1}{c+n} \sum_{i=1}^{n} \delta_{\theta_i}(\cdot)
$$

so $f(x_{n+1} \mid x) \doteq \sum_{j=1}^{MC} f(x_{n+1} \mid \theta^j_{n+1}) / MC$

# Truncated DP

- Recalling the Sethuraman representation, sample from

$$\sum_{h=1}^{K} p_h \delta_{\theta_h}(\cdot)$$

  for sufficiently large $K$ (let $p_K = 1$).
- It's a random finite distribution (Gelfand and Kottas, JCGS 2002)
- Obtain $\theta^j \; : j = 1, ..., MC$ as before
- By conjugacy of DP

$$G \mid \theta = \theta^j \sim DP(c+n, \frac{c}{c+n} G_0(\cdot) + \frac{1}{c+n} \sum_{i=1}^{n} \delta_{\theta_i^j}(\cdot)$$

# Truncated DP

- GK approximate as a truncated DP where the Beta's used to construct $p_h$s are Beta$(1, c+n)$ and the $\theta_h$s are iid from the updated base
- So obtain $\{G_j : j = 1, ..., MC\}$. Inferences about functionals $T(G)$ are based on $\sum_{j=1}^{MC} T(G^j)/MC$
- For example, $T(G) = \int F(x \mid \theta) dG(\theta)$, the CDF for a new observation
- Many contributions including: Doss (1994), Ishwaran and Zarepour (2000), Ishwaran and James (2002), Papaspiliopolus and Roberts (2005), Walker (2007) and Kalli, Griffin and Walker (2009)

# Another Finite Approximation

- Mulliere and Sacci (1995), Ishwaran and Zarepour (2002). Let

$$G_K = \sum_{h=1}^{K} p_h \delta_{\theta_h}(\cdot)$$

with

$$(p_1, ..., p_K) \sim \text{Dirch}(c/K, \ldots, c/K)$$

$$\theta_h \overset{iid}{\sim} G_0$$

- Then for large $K$, $G_K \mathrel{\dot\sim} DP(c, G_0)$

# Finite Approximation

- EXAMPLE: GALAXY DATA. $n = 82$ galaxy velocities obtained from Roeder (1990)
- Approximate a DPM of $N(\mu, 1/\tau)$ variates based on a finite mixture; $K = 50$, $c = 1$
- Take $G_0$ in two dimensions to be the reference prior $N(0, 1000)$ independent of $\text{Gam}(0.001, 0.001)$
- Let $(p_1, \ldots, p_K) \sim \text{Dir}(1/50, \ldots, 1/50)$

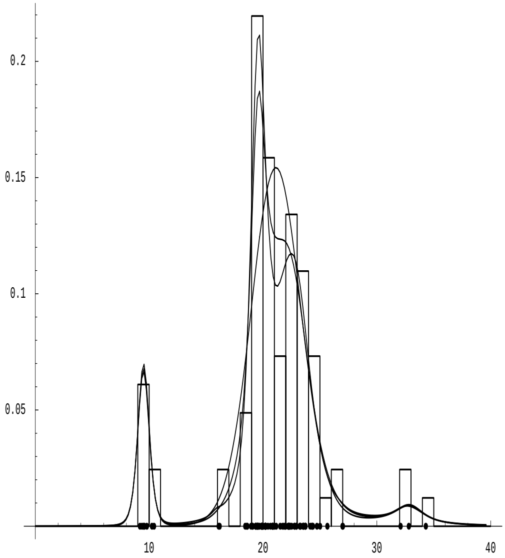



Figure 15.3: *Galaxy data: fits from finite mixture models,* $K = 3, 4, 6$.

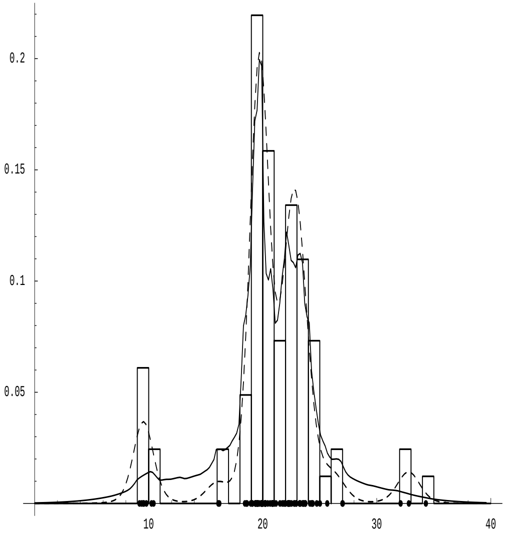



Figure 15.4 *Galaxy data: Dirichlet process mixture (dashed) and mixture of Polya trees (solid) fits.*

# Polya Trees

- Split sample space $\Omega$ into two disjoint sets $B_0$ and $B_1$; further split $B_0$ into $B_{00}$ etc:

| $B_0$ | | $B_1$ | |
|---|---|---|---|
| $B_{00}$ | $B_{01}$ | $B_{10}$ | $B_{11}$ |

-
$$
\begin{aligned}
Y_0 = P(X \in B_0), \qquad Y_1 &= P(X \in B_1), \\
Y_{00} &= P(X \in B_{00} | X \in B_0), \\
Y_{01} &= P(X \in B_{01} | X \in B_0), \\
Y_{10} &= P(X \in B_{10} | X \in B_1), \\
Y_{11} &= P(X \in B_{11} | X \in B_1).
\end{aligned}
$$

- Then $P(X \in B_{ij}) = Y_i Y_{ij}$

# Sets and corresponding conditional probabilities

<table>
<tr><td colspan="8">$\mathbb{R}$</td></tr>
<tr><td colspan="4">$B_0$</td><td colspan="4">$B_1$</td></tr>
<tr><td colspan="8">$(Y_0, Y_1) \sim \text{Dir}(\alpha_0, \alpha_1)$</td></tr>
<tr><td colspan="2">$B_{00}$</td><td colspan="2">$B_{01}$</td><td colspan="2">$B_{10}$</td><td colspan="2">$B_{11}$</td></tr>
<tr><td colspan="4">$(Y_{00}, Y_{01}) \sim \text{Dir}(\alpha_{00}, \alpha_{01})$</td><td colspan="4">$(Y_{10}, Y_{11}) \sim \text{Dir}(\alpha_{10}, \alpha_{11})$</td></tr>
<tr><td>$B_{000}$</td><td>$B_{001}$</td><td>$B_{010}$</td><td>$B_{011}$</td><td>$B_{100}$</td><td>$B_{101}$</td><td>$B_{110}$</td><td>$B_{111}$</td></tr>
<tr><td colspan="2">$(Y_{000}, Y_{001}) \sim \text{Dir}(\alpha_{000}, \alpha_{001})$</td><td colspan="2">$(Y_{010}, Y_{011}) \sim \text{Dir}(\alpha_{010}, \alpha_{011})$</td><td colspan="2">$(Y_{100}, Y_{101}) \sim \text{Dir}(\alpha_{100}, \alpha_{101})$</td><td colspan="2">$(Y_{110}, Y_{111}) \sim \text{Dir}(\alpha_{110}, \alpha_{111})$</td></tr>
</table>

# Probability of partition sets

| $\Omega = [0, 1]$ | | | | | | | |
|---|---|---|---|---|---|---|---|
| $B_0$ | | | | $B_1$ | | | |
| $B_{00}$ | | $B_{01}$ | | $B_{10}$ | | $B_{11}$ | |
| $B_{000}$ | $B_{001}$ | $B_{010}$ | $B_{011}$ | $B_{100}$ | $B_{101}$ | $B_{110}$ | $B_{111}$ |

Instead of $\mathbb{R}$, let's look at $\Omega = [0, 1] \subset \mathbb{R}$.

# Probability of partition sets

| $\Omega = [0, 1]$ | | | | | | | |
|---|---|---|---|---|---|---|---|
| $B_0$ | | | | $B_1$ | | | |
| $B_{00}$ | | $B_{01}$ | | $B_{10}$ | | $B_{11}$ | |
| $B_{000}$ | $B_{001}$ | $B_{010}$ | $B_{011}$ | $B_{100}$ | $B_{101}$ | $B_{110}$ | $B_{111}$ |

Say we want $G(B_{101})$.

# Probability of partition sets

| $\Omega = [0, 1]$ | | | | | | | |
|---|---|---|---|---|---|---|---|
| $B_0$ | | | | $B_1$ | | | |
| $B_{00}$ | | $B_{01}$ | | $B_{10}$ | | $B_{11}$ | |
| $B_{000}$ | $B_{001}$ | $B_{010}$ | $B_{011}$ | $B_{100}$ | $B_{101}$ | $B_{110}$ | $B_{111}$ |

$B_{101} \subset B_{10}$.

# Probability of partition sets

<table>
<tr><td colspan="8">$\Omega = [0, 1]$</td></tr>
<tr><td colspan="4">$B_0$</td><td colspan="4">$B_1$</td></tr>
<tr><td colspan="2">$B_{00}$</td><td colspan="2">$B_{01}$</td><td colspan="2">$B_{10}$</td><td colspan="2">$B_{11}$</td></tr>
<tr><td>$B_{000}$</td><td>$B_{001}$</td><td>$B_{010}$</td><td>$B_{011}$</td><td>$B_{100}$</td><td>$B_{101}$</td><td>$B_{110}$</td><td>$B_{111}$</td></tr>
</table>

$B_{101} \subset B_{10} \subset B_1$.

# Probability of partition sets

<table>
<tr><td colspan="8" align="center">$\Omega = [0,1]$</td></tr>
<tr><td colspan="4" align="center">$B_0$</td><td colspan="4" align="center">$B_1$</td></tr>
<tr><td colspan="2" align="center">$B_{00}$</td><td colspan="2" align="center">$B_{01}$</td><td colspan="2" align="center">$B_{10}$</td><td colspan="2" align="center">$B_{11}$</td></tr>
<tr><td>$B_{000}$</td><td>$B_{001}$</td><td>$B_{010}$</td><td>$B_{011}$</td><td>$B_{100}$</td><td>$B_{101}$</td><td>$B_{110}$</td><td>$B_{111}$</td></tr>
</table>

$$
\begin{aligned}
G(B_{101}) &= G(B_{101} \cap B_{10} \cap B_1) \\
&= G(B_{101}|B_{10}, B_1)G(B_{10}|B_1)G(B_1) \\
&= Y_{101} Y_{10} Y_1 .
\end{aligned}
$$

# $G$-measure of first few sets in $\Pi$

$$
\begin{aligned}
G(B_0) &= Y_0 \\
G(B_1) &= Y_1 \\
G(B_{00}) &= Y_0 Y_{00} \\
G(B_{01}) &= Y_0 Y_{01} \\
G(B_{10}) &= Y_1 Y_{10} \\
G(B_{11}) &= Y_1 Y_{11} \\
G(B_{000}) &= Y_0 Y_{00} Y_{000} \\
G(B_{001}) &= Y_0 Y_{00} Y_{001} \\
G(B_{010}) &= Y_0 Y_{01} Y_{010} \\
G(B_{011}) &= Y_0 Y_{01} Y_{011} \\
G(B_{100}) &= Y_1 Y_{10} Y_{100} \\
G(B_{101}) &= Y_1 Y_{10} Y_{101} \\
G(B_{110}) &= Y_1 Y_{11} Y_{110} \\
G(B_{111}) &= Y_1 Y_{11} Y_{111}
\end{aligned}
$$

- Let $\epsilon = \epsilon_1 \cdots \epsilon_m$ be an arbitrary binary number of dimension $m$
- Split $B_\epsilon \to \{B_{\epsilon 0}, B_{\epsilon 1}\} \qquad \forall \epsilon$.
- Then

$$\left.\begin{array}{l} Y_{\epsilon 0} = P(X \in B_{\epsilon 0} | X \in B_\epsilon) \\ Y_{\epsilon 1} = P(X \in B_{\epsilon 1} | X \in B_\epsilon) \end{array}\right\} \Rightarrow$$

$$P(X \in B_{\epsilon_1 \cdots \epsilon_m}) = \prod_{j=1}^{m} Y_{\epsilon_1 \cdots \epsilon_j}$$

- Random PM for $G$:

$$(Y_{\epsilon 0}, Y_{\epsilon 1}) \sim \text{Beta}(\alpha_{\epsilon 0}, \alpha_{\epsilon 1})$$

- Center on $G_0$ by selecting the partition sets to be appropriate quantiles of $G_0$
- Let $\alpha_\epsilon = cm^2$ at level $m, \forall m$ ($\Rightarrow$ abs cont $G$ w/ prob 1)
- We say $G|G_0, c \sim PT(c, G_0)$, $\qquad E(G(\cdot)) = G_0(\cdot)$
- Finite Polya Tree is truncated at say level $M$
- Large $c$ results in a parametric analysis, and small $c$ results in a more non-parametric analysis

- Partitions defining the Polya tree are induced by *single fixed* centering distribution.
- Sensible choice of $M$ : $2^M \doteq n$
- Will be difficult in practice to specify a single centering distribution.
- Random densities $g(x) = G'(x)$ are discontinuous at every partition point. Infinite number of discontinuities!

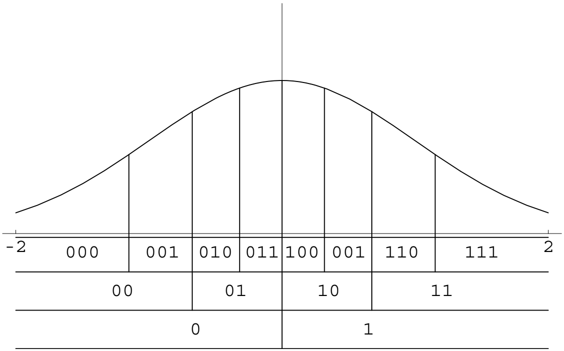



Figure: Finite Polya tree partition sets determined by $G_\theta$:
$\pi_1 = \{B_0, B_1\}$, $\pi_2 = \{B_{00}, B_{01}, B_{10}, B_{11}\}$,
$\pi_3 = \{B_{000}, B_{001}, B_{010}, B_{011}, B_{100}, B_{101}, B_{110}, B_{111}\}$. $G_0 = N(0, 1)$

# Mixture of Finite PTs

- Center on parametric family $\{G_\theta,\ \theta \in \Theta\}$ eg. want
$$E[G(\cdot) \mid \theta)] = G_\theta(\cdot) \quad \forall\theta$$

- Mixtures of Polya trees (Lavine, 1992; Hanson and Johnson, 2002) smooth out partitioning effects and allow robustness against misspecification of (only one) centering distribution
- Prior on $\theta$, $p(\theta)$
- We say $G|G_\theta, c \sim PT(c, G_\theta)$
$$G \sim \int PT(c, G_\theta)p(d\theta)$$
- Predictive density $g(y_{n+1}|Y_1, \ldots, Y_n)$ can be differentiable in infinite tree; random densities $g(y|Y_1, \ldots, Y_n)$ continuous.
- Truncated at level $M$ results in an MFPT
- Large $c$ results in analysis based on the parametric family

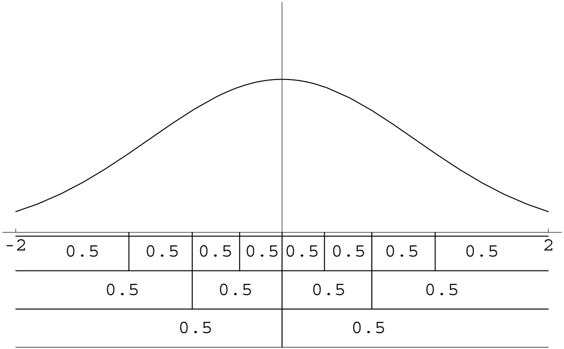



Figure: All pairs $(Y_{\epsilon 0}, Y_{\epsilon 1})$ are 0.5.

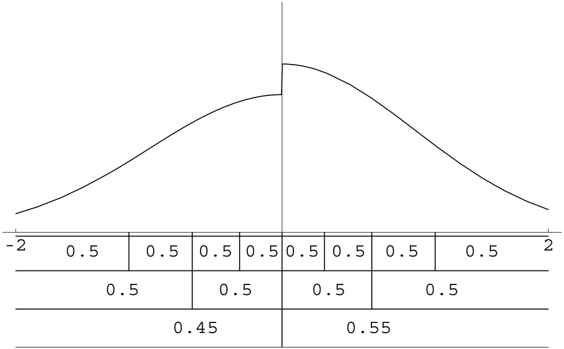



Figure: Pair of level $j = 1$ probabilities $(Y_0, Y_1)$.

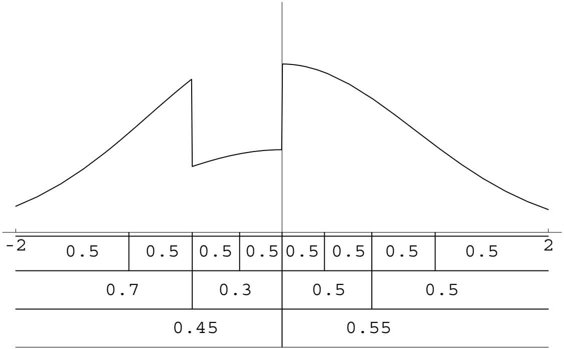



Figure: Pair of level $j = 2$ probabilities $(Y_{00}, Y_{01})$.

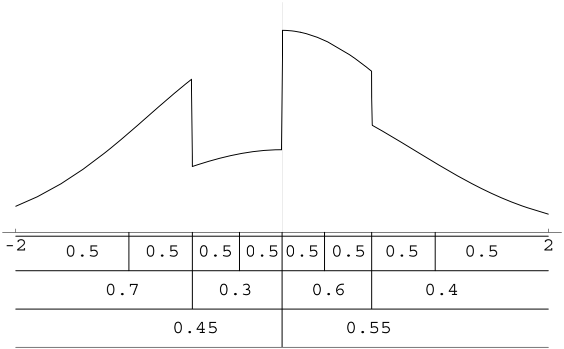



Figure: Pair of level $M = 2$ probabilities $(Y_{10}, Y_{11})$.

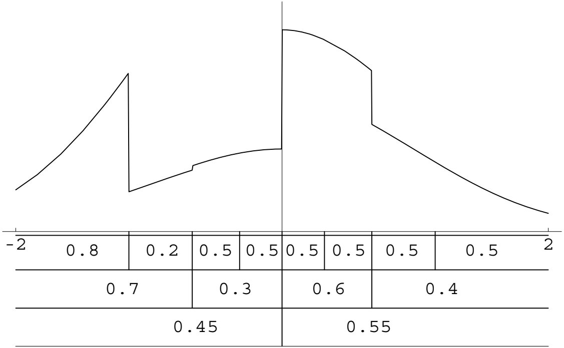



Figure: Pair of level $M = 3$ probabilities $(Y_{000}, Y_{001})$.

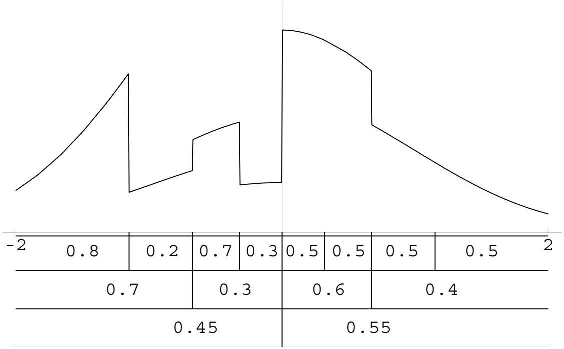



Figure: Pair of level $M = 3$ probabilities $(Y_{010}, Y_{011})$.

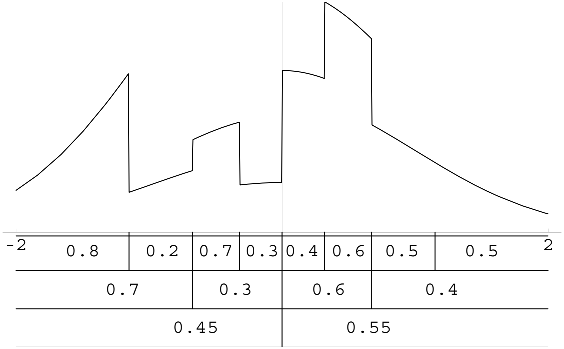



Figure: Pair of level $M = 3$ probabilities $(Y_{100}, Y_{101})$.

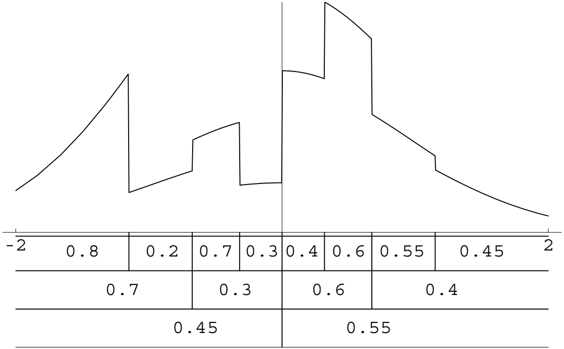



Figure: Pair of level $M = 3$ probabilities $(Y_{110}, Y_{111})$.

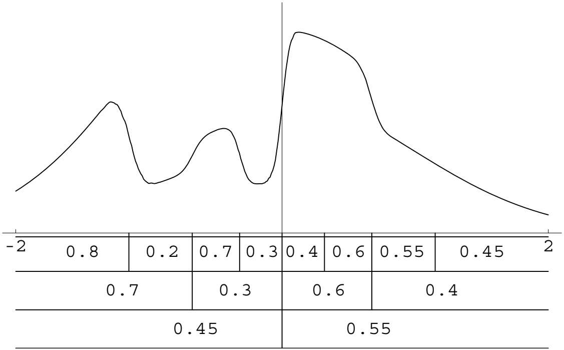



Figure: Mixture of Finite Polya trees.

# Mixture of Finite PTs

- Even with $M = 3$ can get interesting density shapes.

- Allowing $\theta$ to be random smooths density

- Notation: $G \sim PT_M(c, G_\theta)$. $G$ is random probability measure centered at $G_\theta$, parametric on $\mathbb{R}$.

- Further taking $\theta \sim p(\theta)$ induces MFPT.

- $c$ is overall weight attached to $\{G_\theta : \theta \in \mathbf{\Theta}\}$.

# Smoothness properties

**Proposition**

*(Hanson and Johnson, 2002). Let* $G \sim PT_\infty(c, j^2, \Phi_{\mu,\sigma})$ *and* $w_1, \ldots, w_n | G \overset{iid}{\sim} G$*. Let* $(\mu, \sigma^{-2}) \sim N(m, s^2) \times \Gamma(a, b)$*. Then the density of* $g(w_{n+1} | \mathbf{w}_{1:n})$ *is differentiable on* $\mathbb{R} \backslash \{w_1, \ldots, w_n\}$ *but continuous everywhere.*

This also holds for finite MPTs.

**Proposition**

*(Hanson, 2006). Let* $G \sim PT_J(c, j^2, \Phi_{\mu,\sigma})$ *and* $w_1, \ldots, w_n | G \overset{iid}{\sim} G$*. Let* $(\mu, \sigma^{-2}) \sim N(m, s^2) \times \Gamma(a, b)$*. Then the density* $g(w | \mathbf{w}_{1:n}, \mathcal{Y}) = \int_{\boldsymbol{\Theta}} g(w | \mathbf{w}_{1:n}, \mathcal{Y}, \boldsymbol{\theta}) db\theta$ *is differentiable on* $\mathbb{R}$*.*

Holds for multivariate Polya trees as well.

# Random draws PT and MPT

- Simple Polya tree prior $G \sim PT_5(1, \exp(1))$.
- MPT prior $G \sim \int PT_5(1, \exp(\theta))P(d\theta)$ where $\theta \sim \Gamma(10, 10)$ so $E(\theta) = 1$.
- For both $\rho(j) = j^2$, $m = 5$, and $c = 1$.
- Look at densities from 10 random $G$'s.

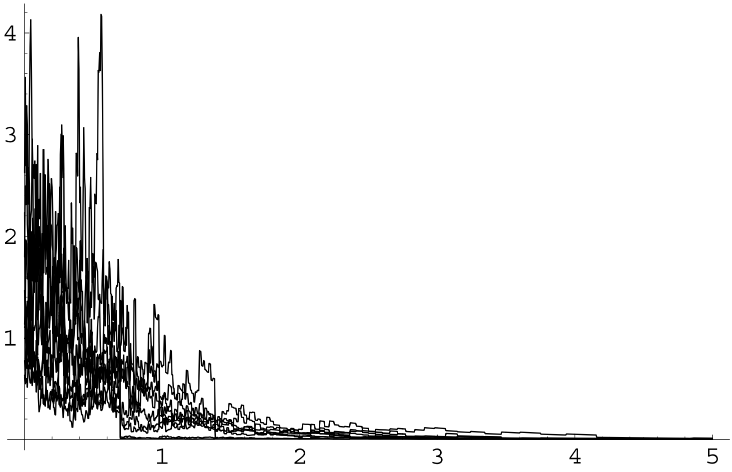



Figure: $G_1, \ldots, G_{10} \overset{iid}{\sim} PT_5(1, \exp(1))$.

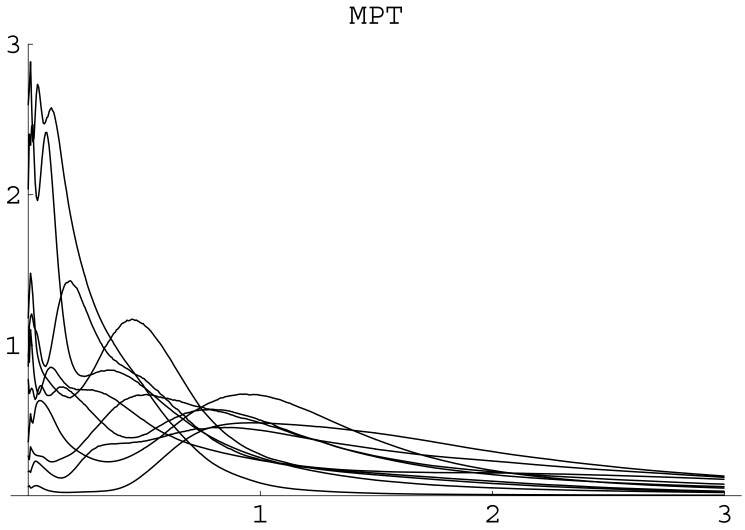



Figure: $G_1, \ldots, G_{10} \overset{iid}{\sim} \int PT_5(1, \exp(\theta))P(d\theta)$.

# Blood Pressure Data

- A randomized study was conducted to assess the assoc between amount of calcium intake and reduction of syst blood pressure (SBP) in black males
- Of 21 healthy black men, 10 were randomly assigned to receive a calcium supplement (group 1) over a 12 week period. The other men received a placebo (group 2)
- The response variable was amount of decrease in systolic blood pressure Negative responses correspond to increases in SBP.
- The data were fitted to the DP, MDP, DPM, PT, and MPT models.

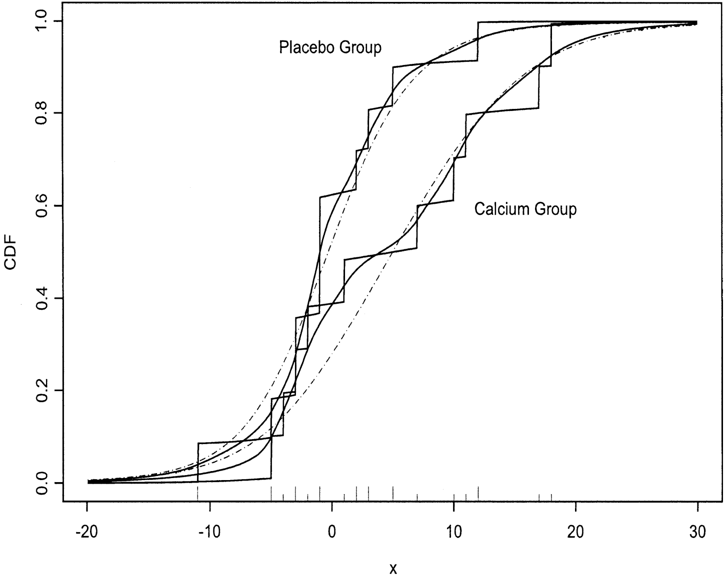



Fig. 2. Blood pressure data: posterior CDF estimates for both groups using the MDP (jagged), DPM (dashed), and MPT (solid) models. The longer tick marks along the $x$-axis correspond to the observed data for the placebo group and the shorter tick marks to the observed data for the calcium group.

Table 1
Blood pressure data: summary statistics for the decrease in systolic blood pressure data for the calcium and placebo groups

| | $n$ | Mean | Median | Std. Dev. | Min | Max |
|---|---|---|---|---|---|---|
| Calcium | 10 | 5.0 | 4 | 8.7 | −5 | 18 |
| Placebo | 11 | −0.27 | −1 | 5.9 | −11 | 12 |

Table 2
Blood pressure data: prior and posterior medians and 95% probability intervals for functionals $T(F)$ for the two-sample problem. The mean and median functionals are denoted by $\mu(\cdot)$ and $\eta(\cdot)$, respectively

| | DP | | MDP | | DPM | |
|---|---|---|---|---|---|---|
| $T(F)$ | Prior | Posterior | Prior | Posterior | Prior | Posterior |
| $\mu(F_1)$ | 5.08<br>(−10.4, 20.3) | 4.96<br>(0.5, 9.9) | 4.90<br>(−14.7, 25.9) | 4.97<br>(0.6, 10.0) | 5.05<br>(−5.2, 16.5) | 5.08<br>(0.3, 9.9) |
| $\mu(F_2)$ | −0.08<br>(−9.5, 9.3) | −0.31<br>(−3.3, 3.0) | 0.02<br>(−16.2, 15.3) | −0.25<br>(−3.2, 3.1) | 0.13<br>(−8.8, 9.6) | −0.30<br>(−3.3, 2.8) |
| $\eta(F_1)$ | 5.01<br>(−10.3, 20.3) | 5.17<br>(−3.0, 11.0) | 4.93<br>(−16.4, 27.6) | 5.27<br>(−3.0, 11.0) | 5.14<br>(−4.6, 15.9) | 4.89<br>(0.2, 9.9) |
| $\eta(F_2)$ | −0.10<br>(−12.4, 11.9) | −1.1<br>(−3.1, 2.9) | −0.10<br>(−17.8, 17.1) | −1.1<br>(−3.1, 2.9) | 0.25<br>(−8.1, 8.7) | −0.35<br>(−3.3, 2.6) |
| $\mu(F_1) - \mu(F_2)$ | 5.12<br>(−9.8, 20.5) | 5.23<br>(−0.3, 11.1) | 4.86<br>(−19.4, 31.5) | 5.23<br>(−0.3, 10.8) | 5.08<br>(−8.94, 20.4) | 5.24<br>(0.0, 10.6) |
| $\eta(F_1) - \eta(F_2)$ | 5.19<br>(−14.0, 24.7) | 4.91<br>(−3.9, 14.1) | 4.86<br>(−22.3, 34.2) | 5.01<br>(−3.9, 14.1) | 5.22<br>(−8.4, 18.9) | 4.99<br>(−0.3, 10.8) |

The DPM model used was, for $k = 1, 2$, and $i = 1, \ldots, n_k$ with $n_1 = 10$, $n_2 = 11$,

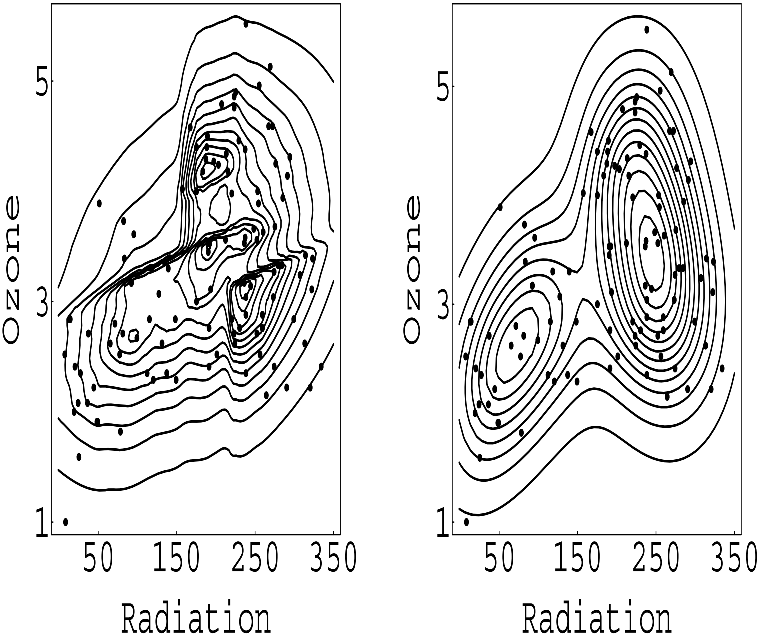
$$x_{ki} | (\mu_{ki}, \sigma^2_{ki}) \overset{\text{ind}}{\sim} N(\mu_{ki}, \tau\sigma^2_{ki})$$

$$(\mu_{ki}, \sigma^2_{ki}) | G_k \overset{\text{ind}}{\sim} G_k$$

$$G_k | \alpha, G_{k0} \overset{\text{ind}}{\sim} \text{DP}(\alpha G_{k0}).$$

# Illustration: Environmental Data

- Bivariate Density Estimation.
- $n = 111$ bivariate observations $w_i = (w_{i1}, w_{i2})'$ on cube root of ozone concentration ($w_{i2}$) and radiation ($w_{i1}$) modeled.
- Previously modeled using DPM of bivariate Gaussian densities.
- Here look at $G \sim \int PT_4(1, \Phi_\theta)dP(\theta)$, where $p(\theta)$ Jeffreys' prior for MVN.
- $BF \approx 45$ in favor of MPT model over Gaussian model.
- MPT model can adapt locally and capture interesting aspects of the data without resorting to finite mixtures...



Ozone
5
3
1
50
150
250
350
Radiation
Ozone
5
3
1
50
150
250
350
Radiation

# Median Regression (Hanson and Johnson, 2002)

- $y_i = x_i\beta + \varepsilon_i \qquad \varepsilon_i \mid G \sim G$

$$G \sim \int FPT_K(c, G_\theta)p(\theta)d\theta$$

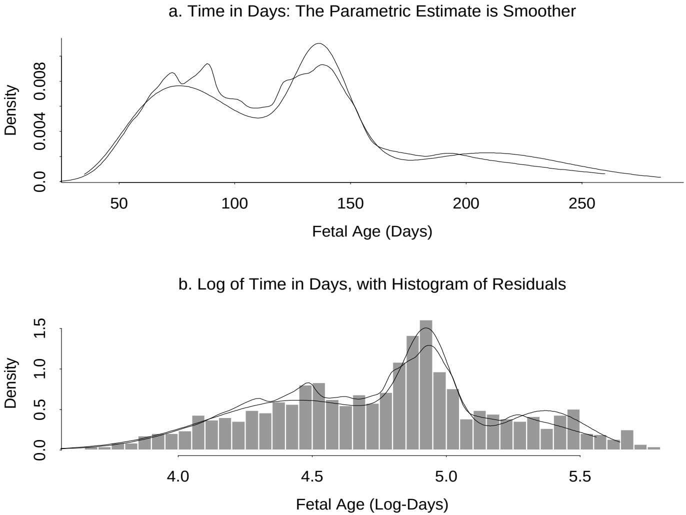
- Errors forced to have median 0, so it's a median regression model, eg.

$$\text{med}(y \mid x) = x\beta$$

# Cow Abortion Data

- Joint modeling of cow-abortion yes/no and time to abortion given that the pregnancy ends in abortion.
- Multiple cycles so need random effects for abortion indicator and time to abortion
- Covariates are NPA (number of prev abortions), Age, Timing of Previous Abortion (early, late, none), Days Open (DO) and Gravidity (Gr)
- Two known causes of abortion:
  - Abortions due to uterine damage (early term abortions)
  - Abortions due to infection (late abortions)

Fig 2. Estimated Parametric and Semi-Parametric Baseline Density of FLD



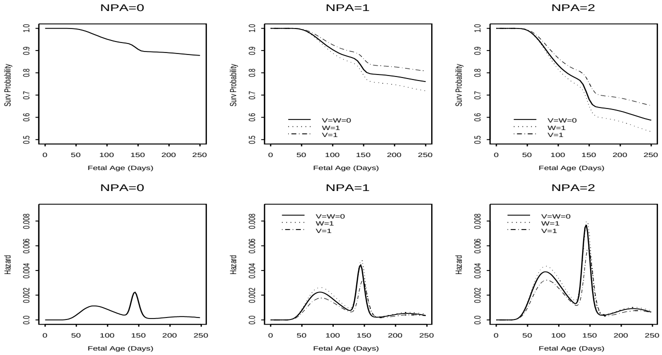

NPA=0
NPA=1
NPA=2
Surv Probability
Hazard
Fetal Age (Days)
V=W=0
W=1
V=1

Table II. Posterior summaries for baseline distribution and variance components.

| | Effect | Mean | Standard deviation |
|---|---|---|---|
| Baseline | $\gamma_1$ | 0.534 | 0.034 |
| | $\gamma_2$ | 0.285 | 0.026 |
| | $\gamma_3$ | 0.181 | 0.016 |
| | $\mu_1$ | 4.453 | 0.068 |
| | $\mu_2$ | 4.924 | 0.055 |
| | $\mu_3$ | 5.374 | 0.058 |
| | $\sigma_1^{-2}$ | 8.809 | 1.077 |
| | $\sigma_2^{-2}$ | 182.6 | 32.94 |
| | $\sigma_3^{-2}$ | 47.96 | 7.503 |
| Variance components | $\lambda_{11}$ | 78.04 | 36.81 |
| | $\lambda_{12}$ | −0.801 | 12.88 |
| | $\lambda_{22}$ | 18.29 | 9.845 |

Table III. Predictive probability of abortion – Logistic model estimates for herds.

| DO | GR | AGE | AB | Herd 3 | Herd 6 |
|---|---|---|---|---|---|
| 40 | 2 | 3 | 0 | 0.143 | 0.077 |
| 40 | 2 | 3 | 1 | 0.293 | 0.173 |
| 40 | 3 | 3 | 0 | 0.107 | 0.056 |
| 40 | 3 | 3 | 1 | 0.228 | 0.129 |
| 150 | 2 | 3 | 0 | 0.140 | 0.075 |
| 150 | 2 | 3 | 1 | 0.287 | 0.168 |
| 150 | 3 | 3 | 0 | 0.096 | 0.050 |
| 150 | 3 | 3 | 1 | 0.207 | 0.116 |
| 40 | 2 | 4.5 | 0 | 0.240 | 0.137 |
| 40 | 2 | 4.5 | 1 | 0.395 | 0.249 |
| 40 | 3 | 4.5 | 0 | 0.183 | 0.101 |
| 40 | 3 | 4.5 | 1 | 0.318 | 0.190 |
| 150 | 2 | 4.5 | 0 | 0.234 | 0.133 |
| 150 | 2 | 4.5 | 1 | 0.388 | 0.243 |
| 150 | 3 | 4.5 | 0 | 0.165 | 0.090 |
| 150 | 3 | 4.5 | 1 | 0.292 | 0.172 |

# Linear Dependent DPM (LDDP)

- MacEachern (1999, 2000), DeIorio et al. 2004, DeIorio et al. 2009
- $$\begin{aligned} f(y_i \mid x_i) &= \int N(y_i \mid x_i\beta, 1/\tau) dG(\beta, \tau) \\ &= \sum_h p_h N(y_i \mid x_i\beta_h, 1/\tau_h) \end{aligned}$$

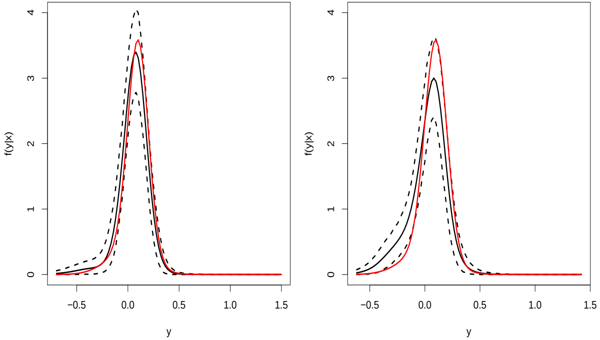
  where $G \sim DP(c, G_\delta)$ and $\delta \sim p(\delta)$
- In the linear case, it's just a DPM of Normal regressions

# Weighted Dependent DPM (WDDP)

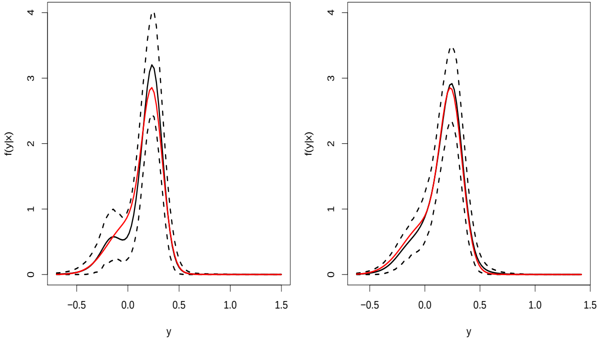
- Müller, Erkanli and West (1996), MacEachern (1999), Griffin and Steel (2006), Dunson, Pillai and Park (2007), Dunson and Park (2008)
- $$\begin{aligned} f(y_i \mid x_i) &= \int N(y_i \mid x_i\beta, 1/\tau) dG_{x_i}(\beta, \tau) \\ &= \sum_h p_h(x_i) N(y_i \mid x_i\beta_h, 1/\tau_h) \end{aligned}$$

  where $p_h(x_i)$ are selected in various clever ways

# DDP results - $x = 0.10$



# DDP results - $x = 0.25$



# DDP results - $x = 0.48$