

**Bayesian Semiparametric Modeling Based on Mixtures of  
Polya Trees**

**Timothy Hanson**

Division of Biostatistics

University of Minnesota

August, 2007

## Overview:

1. Polya tree prior
2. Models for stochastic order
3. Comparison of survival models with same baseline  $S_0$
4. GLMMs

## Polya tree prior

- Notation:  $G \sim PT(c, \rho(\cdot), G_{\theta})$ .  $G$  is random probability measure centered at  $G_{\theta}$ , parametric on  $\mathbb{R}^d$ .
- Further taking  $\theta \sim p(\theta)$  induces MPT.
- Polya tree prior is tail-free (Freedman, 1963) and generalizes the Dirichlet process, where  $\rho(j) = 2^{-dj}$ .
- History of Polya tree dates to 60's & 70's. Early work summarized in Ferguson (1974).
- Mauldin, Sudderth, & Williams (1992) and Lavine (1992, 1994) develop more theory.
- Walker and Mallick (1997, 1999), Walker et al. (1999) use Polya trees in GLMM and survival models. Inference via MCMC.

Aspects of prior construction without detail:

- Polya tree prior on  $G$  defined through nested partitions of  $\mathbb{R}^d$ , say  $\Pi_j^\theta$ , and associated conditional probabilities  $\mathcal{Y}_j$  at level  $j$ .
- Partition  $\Pi_j^\theta$  at level  $j$  splits  $\mathbb{R}^d$  into  $2^{jd}$  pieces.
- $j = 1, 2, \dots, J$ . Rule of thumb:  $J \leq \log_{2^d} n$ .
- At level  $J$ , let  $G$  follow  $G_\theta$ .
- Have  $E\{G(A)\} = G_\theta(A)$ .  $\text{var}\{G(A)\}$  depends on overall weight  $c$  and function  $\rho(\cdot)$ .
- Nice feature of Polya tree:  $G$  can have a density w.r.t. Lebesgue measure when  $J = \infty$ ; need  $\sum_{j=1}^{\infty} \frac{1}{\rho(j)} < \infty$ .

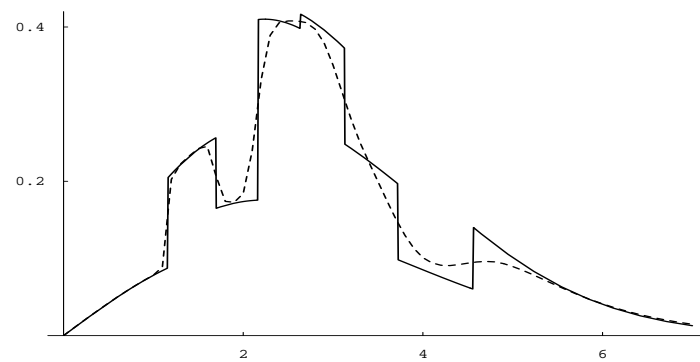
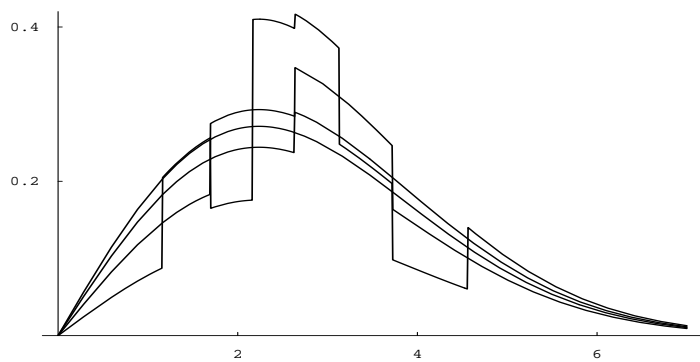
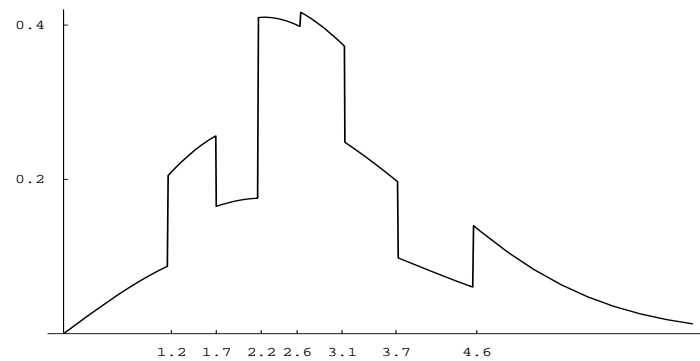
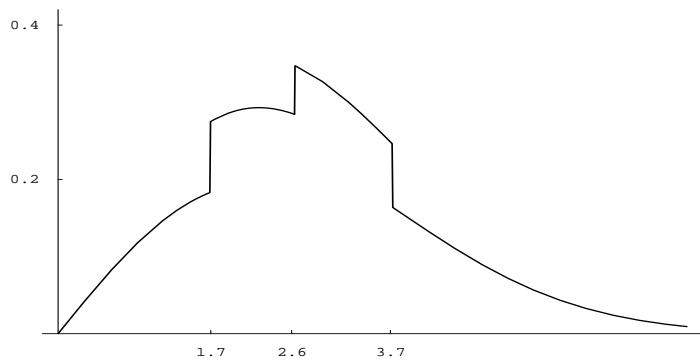
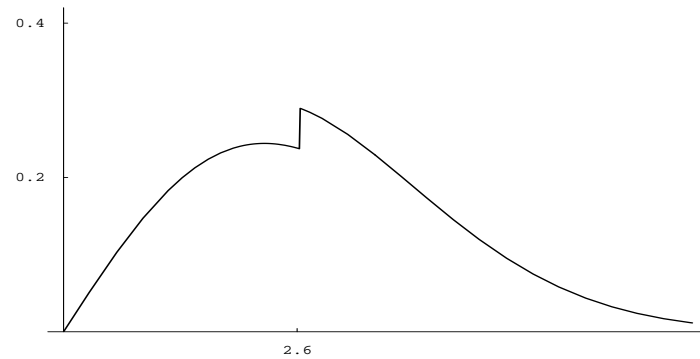


Figure 1:  $J = 0, 1, 2, 3$  for Weibull(2,10).

Bit more detail:

- $B_{\boldsymbol{\theta}}(j, \mathbf{k}) \in \Pi_j^{\boldsymbol{\theta}}$  chosen so  $G_{\boldsymbol{\theta}}\{B_{\boldsymbol{\theta}}(j, \mathbf{k})\} = \frac{1}{2^{jd}}$ .
- $Y(j, \mathbf{k})$  associated conditional probabilities of an observation being in one of  $2^d$  offspring sets at level  $j + 1$  given the observation is in  $B_{\boldsymbol{\theta}}(j, \mathbf{k})$ .

$$Y(j, \mathbf{k}) \sim \text{Dirichlet}(c\rho(j)\mathbf{1}_{2^d}).$$

- $c$  and  $\rho(j)$  affect how quickly data “take over” parametric centering family  $\{G_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ .
  - $c$  is overall weight attached to  $G_{\boldsymbol{\theta}}$ .  $\rho(j)$  affects how “clumped” data are.  $\rho(j) = j^2$  often used.
  - $\rho_{\tau}(j) = 2^{\tau j}$ ,  $\tau \sim p(\tau)$  can add flexibility.

## Mixtures of Polya trees:

- MPT considered by Lavine (1992), Berger and Guglielmi (2001), Hanson and Johnson (2002), Hanson (2006), & some others.
- MPT can smooth out partitioning effects of PT.
- MCMC can be straightforward to set up based on underlying parametric model, but mixing is poor if underlying model grossly incorrect.
- Ongoing research: adaptive MCMC for MPT models.
- MPT generalizes MDP, but consideration of models with densities  $g(x) = G'(x)$  obviates keeping track of tied data.
- Rest of talk looks at three examples of use of MPTs with some comparison to other nonparametric priors.

## 1. Stochastic order in modeling serology score data

Enzyme-Linked Immunosorbent Assay (ELISA) indirectly measures antibodies specific to an antigen linked to an enzyme which causes a chromogenic or fluorogenic substrate to produce a signal.

$X$  is serologic test result reflecting amount of antibodies present in blood sample.

$G$  = distribution of serology scores from infected population;  $F$  = distribution from noninfected. Reasonable ELISA satisfies stochastic order:

$$F \leq_{st} G \Leftrightarrow G(t) \leq F(t) \text{ for all } t.$$

A constructive approach (Gelfand and Kottas, 2001):

$$F(t) = H_1(t), \quad G(t) = H_1(t)H_2(t),$$

where  $H_1(t)$  and  $H_2(t)$  are CDFs.

Two models:

- MPT,  $H_j(t) \sim PT(c_j, \rho, N(\mu_j, \tau_j^{-1}))$  for  $j = 1, 2$ .
- DPM,  $H_j(t) \sim DP(\alpha_j, N(\mu_j, \tau_j^{-1}))$  for  $j = 1, 2$  and

$$F(t|H_1, H_2, \sigma^2) = \int N(t; x, \sigma^2) dH_1(x),$$

$$G(t|H_1, H_2, \sigma^2) = \iint N(t; \max\{x, y\}, \sigma^2) dH_1(x) dH_2(y)$$

Both models completed by  $\mu_j \sim N(m_j, v_j)$  and  $\tau_j^{-1} \sim \Gamma(a_j, b_j)$ .

DPM also requires  $\sigma^{-2} \sim \Gamma(a_3, b_3)$ .

Matching priors:

Note that given  $(\mu_1, \mu_2, \tau_1, \tau_2)$ ,

- PT  $\Rightarrow E\{F(A)\} = N(A|\mu_1, \tau_1^{-1})$ .
- DPM  $\Rightarrow E\{F(A)\} = N(A|\mu_1, \tau_1^{-1} + \sigma^2)$ .

WLOG set  $\mu_1 = 0$ . For PT also set  $\tau_1 = 1$ . For DPM constrain  $\tau_1^{-1} + \sigma^2 = 1$  so then

- PT  $\Rightarrow E\{F(A)\} = N(A|0, 1)$ .
- DPM  $\Rightarrow E\{F(A)\} = N(A|0, 1)$ .

Predictive densities are matched. Define  $r = \frac{\sigma^2}{\sigma^2 + 1/\tau_1}$ .  $r \in (0, 1)$  measures how spread out the mixture component locations are relative to the variability of each component. When  $r$  is small, prior density draws are more wiggly.

Prior variability about  $N(0, 1)$  measured by  $L_1$ -norm:

$$\|f - N(0, 1)\|_1 = \int_{\mathbb{R}} |f(t) - \phi(t)| dt.$$

- Invariant to location-scale transformation of prior centering distribution for both MPT & DPM models.
- Useful interpretation in terms of total variation norm,

$$\|f - N(0, 1)\|_1 = 2 \left[ \sup_{A \subset \mathbb{R}} |F_0(A) - N(A|0, 1)| \right].$$

- Simulation:  $\alpha_1 \sim \Gamma(2, 2)$  and  $c_1 \sim \Gamma(2, 1.5)$  produce a prior median and 95% CI of 0.5 and (0.2,1.3) for  $\|f_0 - N(0, 1)\|_1$ .
- Same priors used for  $\alpha_2$  and  $c_2$ .

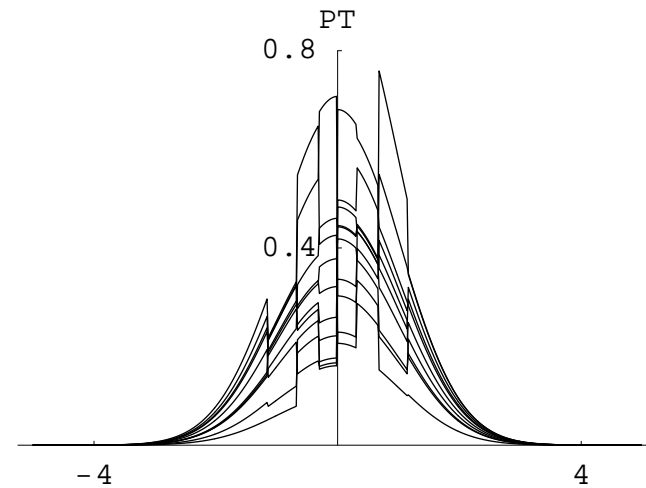
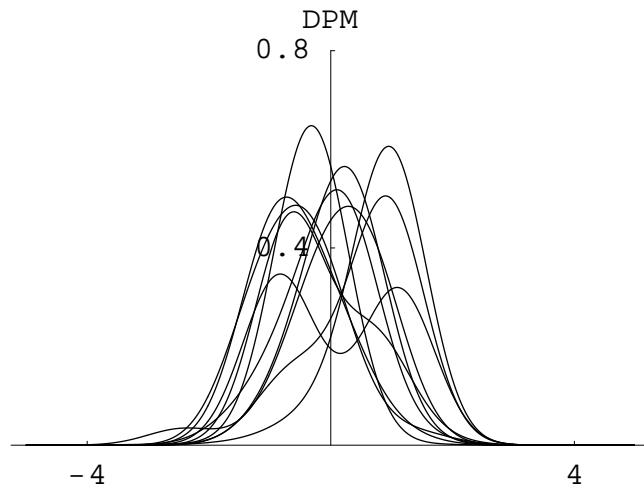


Figure 2: DPM & MPT prior density draws matched on  $L_1$  distance from expected density:  $\|f - N(0, 1)\|_1$ .

## Data analysis:

- Johne's disease in dairy cattle from ELISA developed by Institut Pourquier in Montpellier, France.
- $n_0 = 345$  noninfected and  $n_1 = 258$  infected from 14 herds in Minnesota and Wisconsin.
- $\text{LPML}_0 = -314.5$  for MPT,  $\text{LPML}_0 = -315.9$  for DPM. MPT slightly better in noninfected.
- $\text{LPML}_1 = -369.7$  for MPT,  $\text{LPML}_1 = -365.3$  for DPM. DPM slightly better in infected.

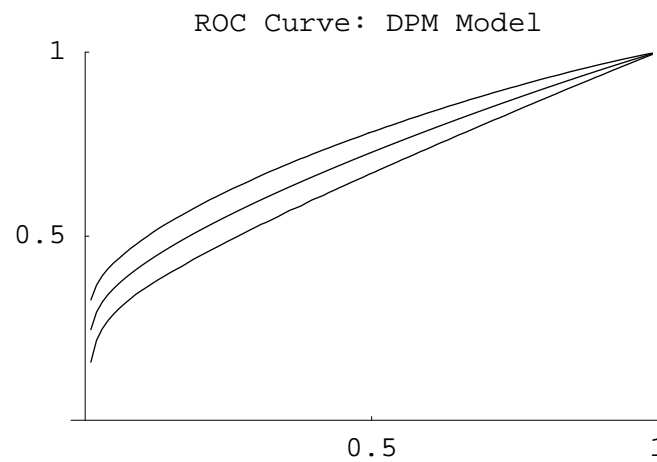
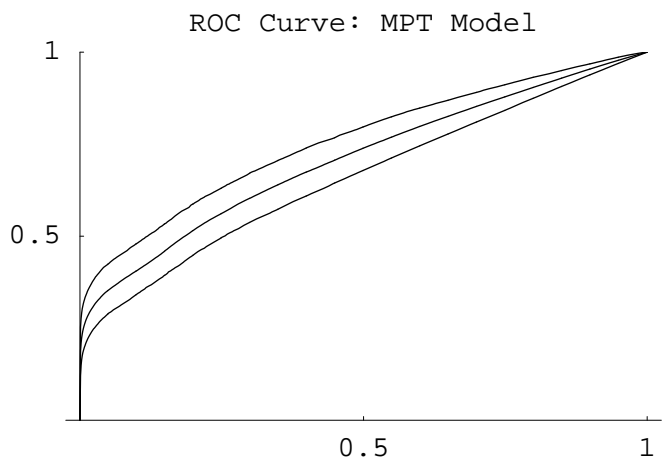
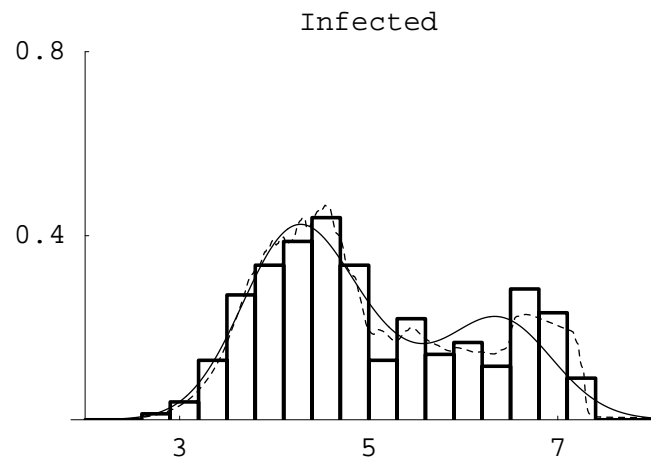
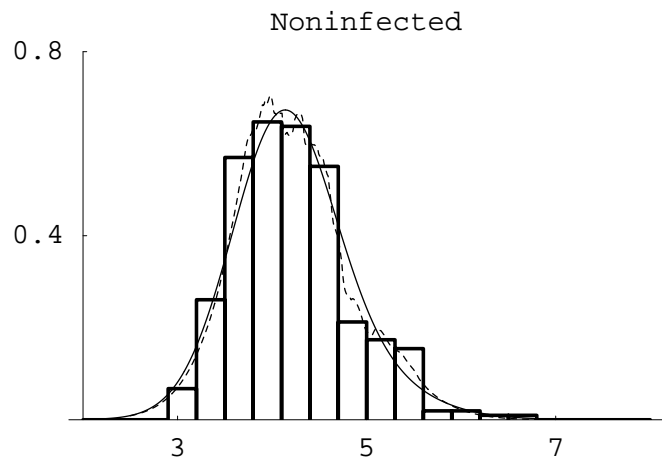


Figure 3: DPM & MPT densities; Institut Pourquoi ELISA.

## 2. Choosing among survival models

SEER data: look at survival of  $j^{\text{th}}$  woman in county  $i = 1, \dots, 99$  from Iowa.  $ij^{\text{th}}$  linear predictor  $\eta_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \gamma_i$ . Examined 3 models:

- Proportional hazards (PH):

$$S_{\mathbf{x}_{ij}}(t) = S_0(t)^{\exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + \gamma_i)},$$

- Accelerated failure time (AFT):

$$S_{\mathbf{x}_{ij}}(t) = S_0\{\exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + \gamma_i)t\},$$

- Proportional odds (PO):

$$\frac{S_{\mathbf{x}_{ij}}(t)}{1 - S_{\mathbf{x}_{ij}}(t)} = \exp(\mathbf{x}'_{ij}\boldsymbol{\beta} + \gamma_i) \frac{S_0(t)}{1 - S_0(t)}.$$

Intrinsic CAR prior or *iid*  $N(0, \tau^{-1})$  on  $\{\gamma_i\}_{i=1}^{99}$ .

$$\text{CAR: } \gamma_i | \gamma_{-i}, \tau \sim N \left( \frac{1}{w_{i\bullet}} \sum_{j:w_{ij}=1} \gamma_j, \frac{1}{\tau w_{i\bullet}} \right),$$

where  $w_{ij} = 1$  if county  $i$  borders county  $j$ , zero otherwise.

**Question:** which is predictively most important?

- Parametric versus nonparametric assumptions on baseline survival  $S_0$ , or...
- assumptions on frailty terms, or...
- assumptions built into survival model (PH, AFT, PO) itself?

Note that frailties enter into linear predictor  $\eta_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \gamma_i$ . If overall model is grossly invalid then no way to “fix” frailty distribution or assumptions on  $S_0$  to make model fit adequate. Need to consider alternative models.

## Analysis of the 1995-1998 Iowa SEER data

- 1073 women in Iowa, who were diagnosed with malignant breast cancer starting in 1995, with enrollment and follow-up continued through the end of 1998.
- Only deaths due to metastasis of cancerous nodes in the breast were considered to be events.
- 488 events; 585 censorings.
- Covariates: race (white or other), age in years at diagnosis, number of primaries, and the stage of the disease: local (baseline, confined to the breast), regional (spread beyond the breast tissue), or distant (metastatis).

## Results for SEER data

model	$c$ prior	PH	AFT	PO
MPT CAR	$\Gamma(5, 1)$	-2224.7	-2236.1	-2208.8
frailty	$\Gamma(20, 2)$	-2229.0	-2235.2	-2211.6
MPT i.i.d.	$\Gamma(5, 1)$	-2233.7	-2252.3	-2218.4
frailty	$\Gamma(20, 2)$	-2239.5	-2252.9	-2220.5
MPT	$\Gamma(5, 1)$	-2239.7	-2234.6	-2223.5
non-frailty	$\Gamma(20, 2)$	-2242.1	-2235.9	-2224.1
non-MPT CAR	—	-2243.1	-2237.0	-2230.3
non-MPT i.i.d.	—	-2253.7	-2257.5	-2242.6
non-MPT non-frailty	—	-2254.9	-2239.5	-2239.3

Table 1: LPML for the competing PH, AFT, and PO models, with the MPT centered at the log-logistic baseline.

- Among survival models, overall PO>PH>AFT.
- Overall, MPT>log-logistic.
- For PO and PH models, CAR>i.i.d.>none.
- For AFT model CAR≈none>i.i.d.
- Overall, survival model most important, followed by assumptions on baseline, *followed by frailty model*.
- Focus in literature is on development of complex frailty models within context of PH; alternative survival models often not considered.

covariates	PH	AFT	PO
centered age	0.018 (0.012, 0.024)	0.017 (0.011, 0.024)	-0.030 (-0.038, -0.021)
regional stage	0.22 (-0.02, 0.45)	0.19 (-0.03, 0.40)	-0.49 (-0.75, -0.21)
distant stage	1.64 (1.41, 1.88)	1.50 (1.23, 1.77)	-2.70 (-3.01, -2.35)

Table 2: Posterior medians and 95% equal-tail credible intervals, MPT CAR frailty model fixed effects. Note “regional” significant under PO model, not others.

- Three models fit using same nonparametric prior on  $S_0$ . Differences attributable to model assumptions, not different priors.
- PO model difficult to fit via DP, DPM, beta, gamma, extended versions of these.
- MCMC scheme based on initial fits of corresponding parametric models.
- Generalizing to dependent Polya tree processes across time or space.

### 3. Generalized linear mixed models

Response  $y_{ij}$  where

$$y_{ij} \sim \text{Poisson}(e^{\eta_{ij}})$$

$$y_{ij} \sim \text{bin} \left( n_{ij}, \frac{e^{\eta_{ij}}}{1 + e^{\eta_{ij}}} \right)$$

$$y_{ij} \sim \text{bin} (n_{ij}, \Phi(\eta_{ij}))$$

$$P(y_{ij} = k) = \Phi(\alpha_k + \eta_{ij}) \text{ for } k = 1, \dots, K$$

$$y_{ij} \sim N(\eta_{ij}, \sigma^2)$$

$$y_{ij} \sim \Gamma(e^{\eta_{ij}}, \nu)$$

DPpackage by Alejandro Jara implements: Poisson, logistic, probit, cumulative probit-link for ordered categorical data, normal, and gamma regression models with random effects. Linear predictor  $\eta_{ij}$  modeled through several Bayes NP priors...

Linear predictor is

$$\eta_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{b}_i, \text{ where } \mathbf{b}_1, \dots, \mathbf{b}_m | G \stackrel{iid}{\sim} G.$$

DPpackage can fit (among *many* other Bayes NP models):

- $G \sim DP(\alpha, G_{\boldsymbol{\theta}})$
- $g(\mathbf{x}) = \int_{\mathbb{R}^d} \phi(\mathbf{x}|\mathbf{m}, \boldsymbol{\Omega}) dH(\mathbf{m})$  where  $H \sim DP(\alpha, G_{\boldsymbol{\theta}})$
- $G \sim PT(c, \rho(\cdot), G_{\boldsymbol{\theta}})$

$G_{\boldsymbol{\theta}} = N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ . Prior placed on  $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta})$ . Also:  $\boldsymbol{\Omega}, \sigma^{-2}, \nu, \boldsymbol{\alpha}$ .

That is, looked at MDP, MDPM, and MPT priors on random effects distribution  $G$ .

## Multivariate MPT

- Developed by Paddock (1999); Paddock et al. (2003); Hanson (2006); Jara, Hanson, and Lesaffre (2007); Hanson, Branscum, and Gardner (2008).
- Many ways to define partition sets  $B_{\theta}(j, \mathbf{k})$ . Natural, computationally tractable approach is to consider  $\Sigma = \mathbf{U}\mathbf{U}'$  where  $\mathbf{U}$  comes from (a) Cholesky, (b)  $\mathbf{U} = \mathbf{M}\mathbf{\Lambda}^{1/2}$ , or (c)  $\mathbf{U} = \mathbf{M}\mathbf{\Lambda}^{1/2}\mathbf{M}'$ .  $\mathbf{M}$  and  $\mathbf{\Lambda}$  come from spectral decomposition. Then consider affine transformation of “canonical” multivariate Polya tree centered at  $N_d(\mathbf{0}, \mathbf{I})$ .
- Too many parameters to sample  $\mathcal{Y}_j$ . Instead, we marginalize and base inference on  $p(\mathbf{b}_1, \dots, \mathbf{b}_m | \boldsymbol{\mu}, \Sigma)$ .

**LMM example:** Carlin & Louis (2000) and Basu & Chib (2003) look at  $\sqrt{\text{CD4}}$  counts  $y_{ij}$  for individual  $i$  at time  $t_j$ ,  $i = 1, \dots, 467$ ,  $j = 1, \dots, n_i$ . The model for each subject is

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\epsilon}_i,$$

where we assume

$$\mathbf{b}_1, \dots, \mathbf{b}_{467} | G \stackrel{iid}{\sim} G, \quad G \sim \int PT_5(c, \rho, \Phi_{\boldsymbol{\Sigma}}) dP(c, \boldsymbol{\Sigma}),$$

and have the usual conjugate priors for  $\boldsymbol{\beta}$  and  $\boldsymbol{\epsilon}$  provided in Carlin and Louis. Prior  $\boldsymbol{\Sigma}^{-1} \sim \text{Wishart}(24, \mathbf{S}_0)$  also provided in C & L.

Long story short: (1) BF for MPT versus parametric model is greater than  $10^{250}$ , (2) BF for MPT versus equivalent DPM model is greater than  $10^{200}$ .

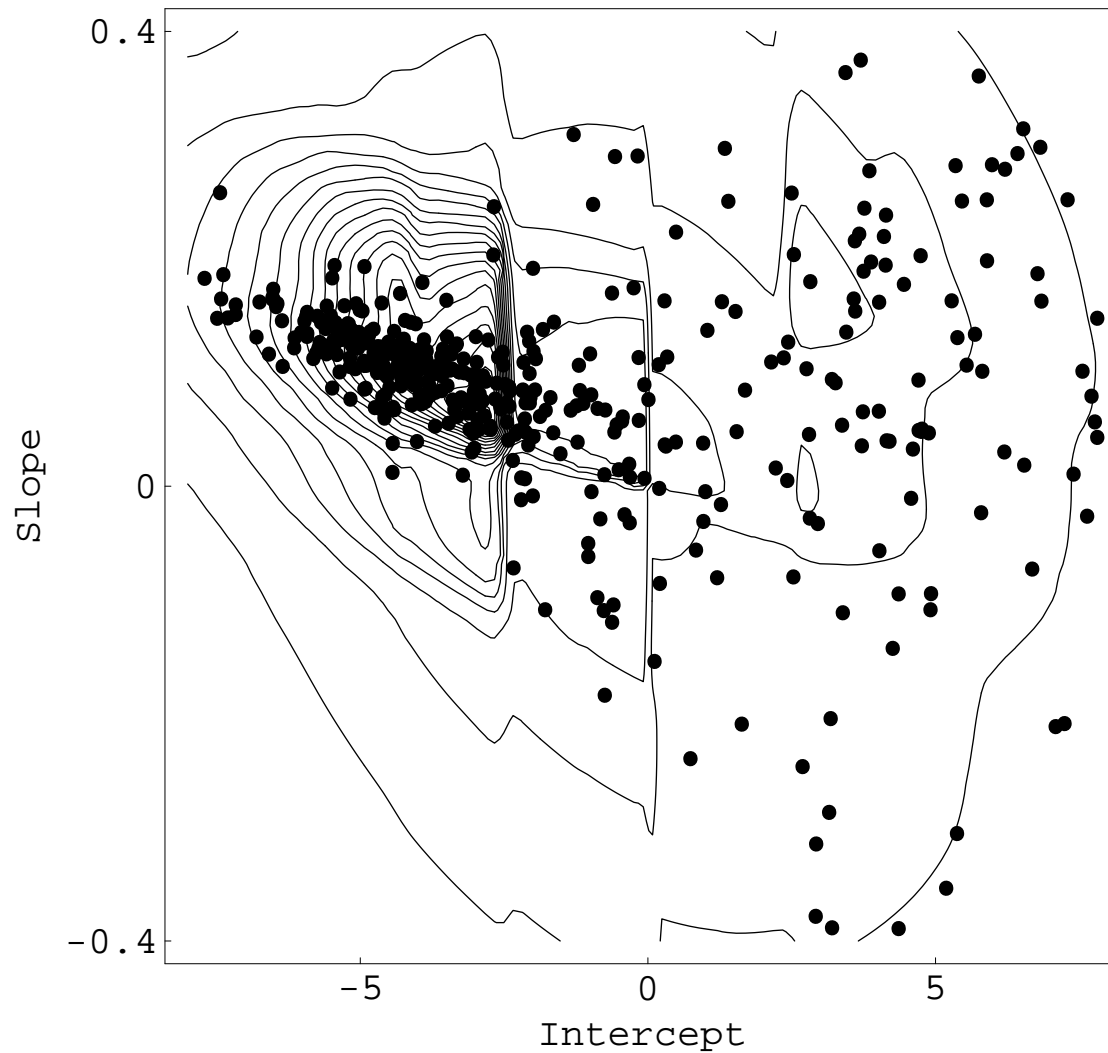


Figure 4: Predictive density level curves of  $G$  with  $E(\mathbf{b}_i | \mathbf{y}_1, \dots, \mathbf{y}_{467})$ .

## Comments...

- MPT have not been tapped to full potential.
- No need to convolve with continuous kernel to get absolutely continuous measure.
- Often fits better than mixture of normals; especially data that exhibit drastic change over small area. Draper (1999) notes “wavelet-like” properties.
- If underlying parametric family okay, not losing much. Can formally test using Bayes factors through Savage-Dickey ratio.
- Collaborators attached to this work: Alejandro Jara, Luping Zhao, Thanasis Kottas, Adam Branscum, Wes Johnson, and Brad Carlin.
- Thanks!