

**A Natural Nonparametric Generalization of Parametric  
Statistical Models**

**Tim Hanson**

Division of Biostatistics

University of Minnesota

November, 2008

- To Bayes or not to Bayes...
- The Polya tree prior.
- Applications: mixed models and survival analysis.
  - Meta analysis.
  - Generalized linear mixed models with exchangeable and partially exchangeable random effects.
  - Multivariate Polya trees.

## Why Bayes?

- Incorporation of existing prior information.
- Through MCMC, able to fit models cannot fit otherwise.
- Complex, multi-level nested and/or crossed GLMMs relatively painless to fit. Approximations at level of likelihood not necessary.

## Why *not* Bayes?

- Subjective? Choosing a model is subjective!
- Computationally intensive.
- Lack of easy to use software? Not anymore: WinBUGS & brugs, DPpackage (and many more) for R, SAS proc mcmc, BayesX, more...

## Bayesian nonparametrics

- Why? One word: flexibility.
- Improved prediction, more realistic models.
- Several priors to choose from: Dirichlet process and other stick breaking processes (Pitman-Yor, species sampling, etc.), normalized inverse Gaussian, beta, gamma, extended (smoothed) versions, Polya trees, Bernstein polynomials, wavelets, penalized splines, etc.
- For distributional modeling, take prior on space of all probability distributions.
- Example: On next slide is 10 *iid* random densities from a mixture of Polya trees (MPT) prior centered at the  $\exp(1)$  distribution...

MPT

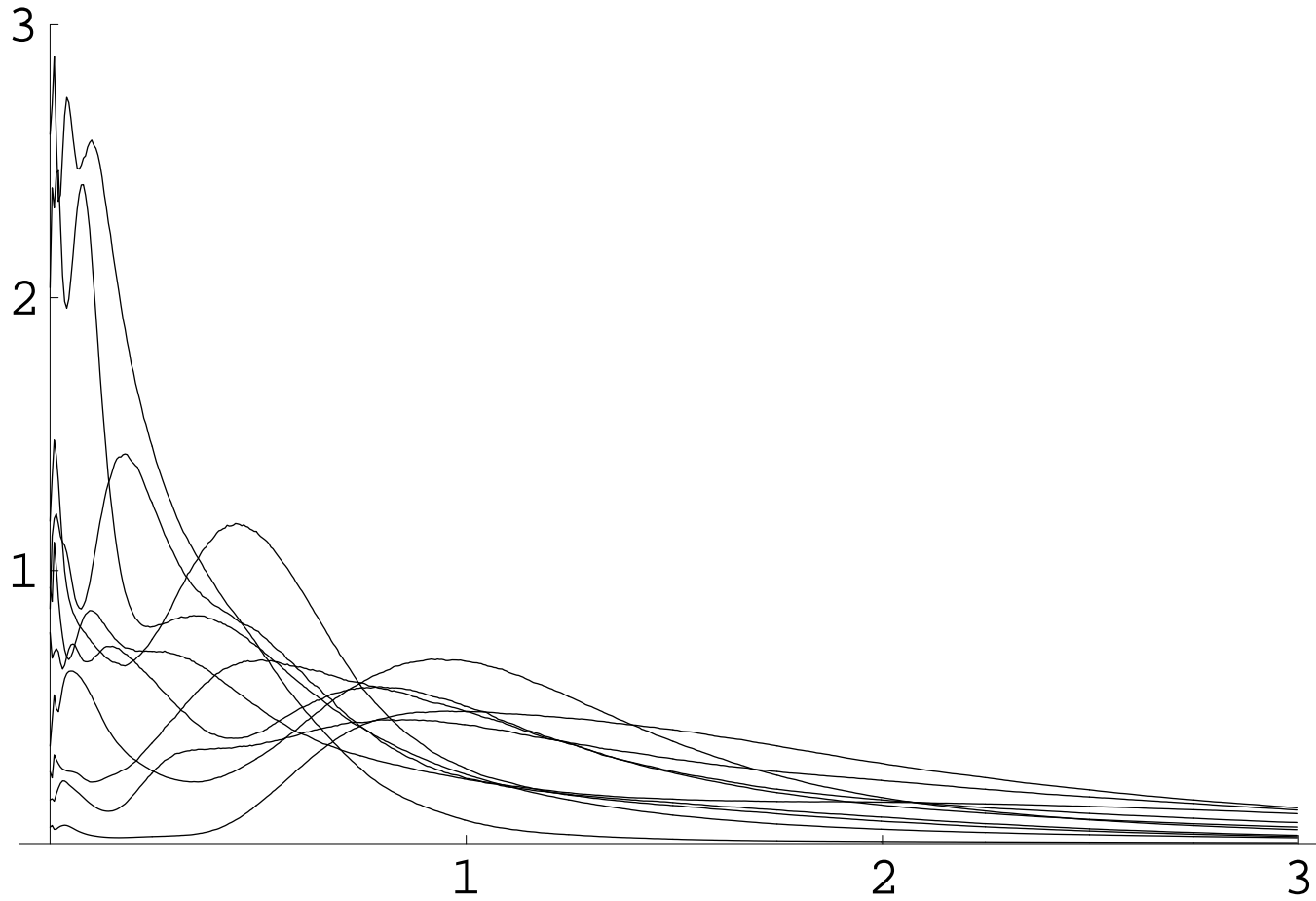


Figure 1:  $g_1, \dots, g_{10} \stackrel{iid}{\sim} \int PT_5(1, \rho, \exp(\theta)) P(d\theta)$ .  $E(\theta) = 1$ .

# Polya trees generalize existing parametric families

- Can add parameters  $\{Y_0, Y_{00}, Y_{10}, Y_{000}, Y_{010}, Y_{100}, Y_{110}\}$  to  $(\mu, \sigma^2)$  to make  $N(\mu, \sigma^2)$  more flexible.

## Univariate Polya tree prior

- Notation:  $G \sim PT(c, \rho(\cdot), G_{\theta})$ .  $G$  is random probability measure centered at  $G_{\theta}$ , parametric on  $\mathbb{R}$ .
- Polya tree prior is tail-free (Freedman, 1963); Dirichlet process special case  $\rho(j) = 2^{-j}$ .
- 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.

## Polya tree = partition + conditional probabilities

- Polya tree prior on  $G$  defined through nested partitions of  $\mathbb{R}$ , say  $\Pi_j^\theta$ , and associated conditional probabilities  $\mathcal{Y}_j$  at level  $j$ .
  - Finite tree has  $j \leq J$ .
  - On level  $J$  sets  $G$  follows  $G_\theta$ .
  - Rule of thumb:  $J \leq \log_2 n$ .
- Partition  $\Pi_j^\theta$  at level  $j$  splits  $\mathbb{R}$  into  $2^j$  pieces of equal probability under  $G_\theta$ . Sets denoted  $B_\theta(\epsilon)$  where  $\epsilon$  is binary.
- Next slide shows  $\Pi_1$ ,  $\Pi_2$ , and  $\Pi_3$  for  $G_\theta = N(0, 1)$ .

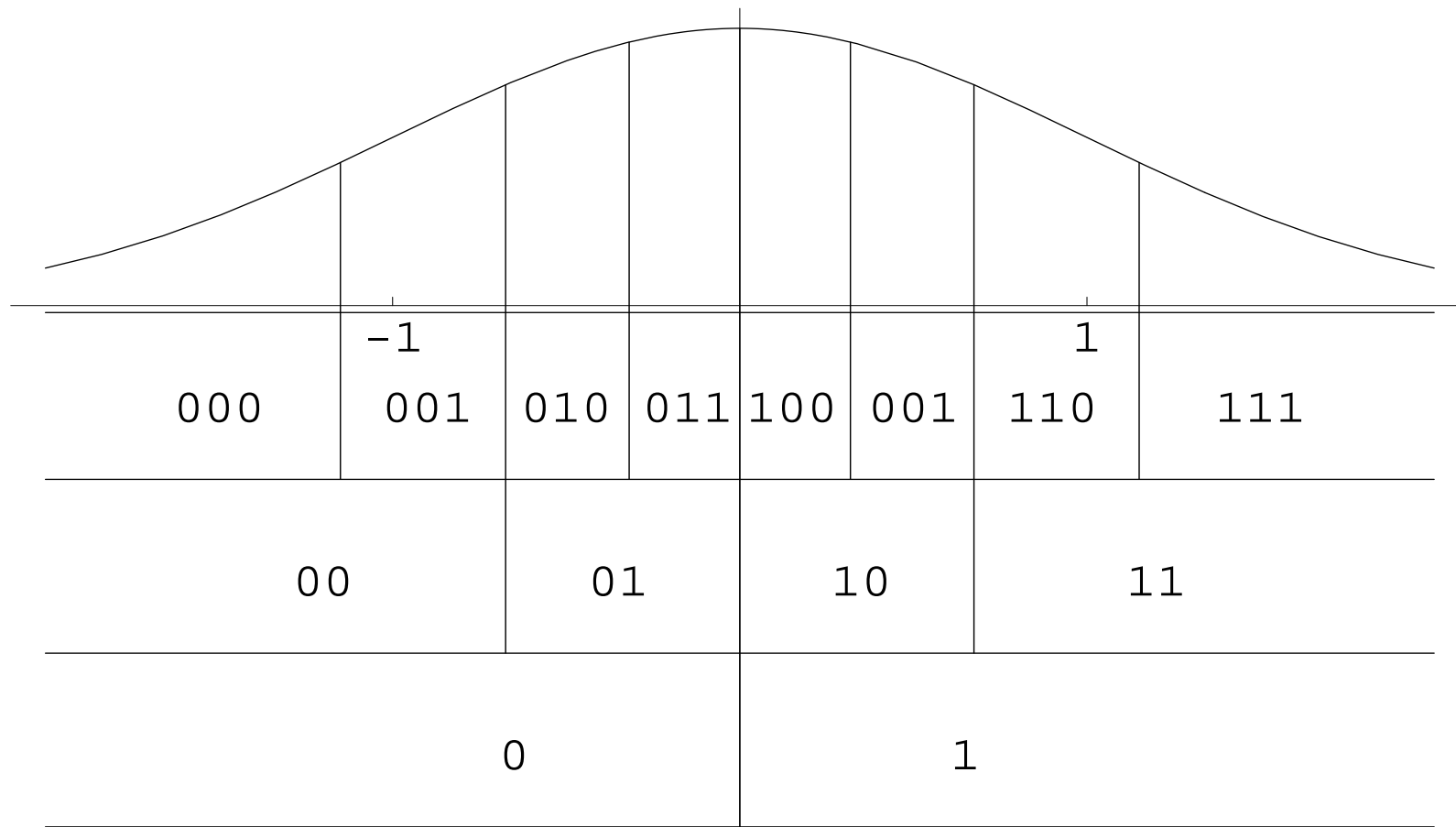


Figure 2: First 3 partitions of  $\mathbb{R}$  generated by  $N(0, 1)$ .

- Parametric  $G_{\theta}$  gives partition.
- Add  $\mathcal{Y}_1 = \{Y_0, Y_1\}$ ,  $\mathcal{Y}_2 = \{Y_{00}, Y_{01}, Y_{10}, Y_{11}\}$ ,  
 $\mathcal{Y}_3 = \{Y_{000}, Y_{001}, Y_{010}, Y_{011}, Y_{100}, Y_{101}, Y_{110}, Y_{111}\}$ , etc. to refine density shape. Let  $\mathcal{Y} = \{\mathcal{Y}_1, \mathcal{Y}_2, \dots, \mathcal{Y}_J\}$ .
- $Y_{\epsilon 0} = G\{B_{\theta}(\epsilon 0) | B_{\theta}(\epsilon)\}$ .  $Y_{\epsilon 1} = 1 - Y_{\epsilon 0}$ .
- Next slide uses  $G_{\theta} = \text{Weibull}(2, 10)$ .
  - Upper left: all  $Y_{\epsilon} = 0.5$  gives back  $\text{Weibull}(2, 10)$ .
  - Upper right:  $(Y_0, Y_1) = (0.45, 0.55)$ .
  - Middle left: adds  $(Y_{00}, Y_{01}) = (0.4, 0.6)$ ,  $(Y_{10}, Y_{11}) = (0.6, 0.4)$ .
  - Middle right:  $Y_{000} = 0.3$ ,  $Y_{010} = 0.3$ ,  $Y_{100} = 0.6$ ,  $Y_{110} = 0.3$ .
  - Lower right: Keeping  $\mathcal{Y}$  as above but mixing over Weibull parameters.

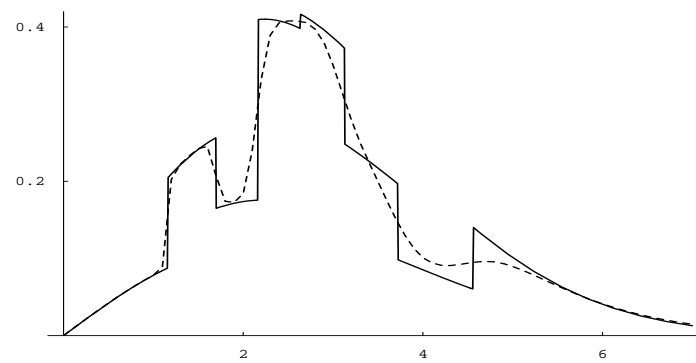
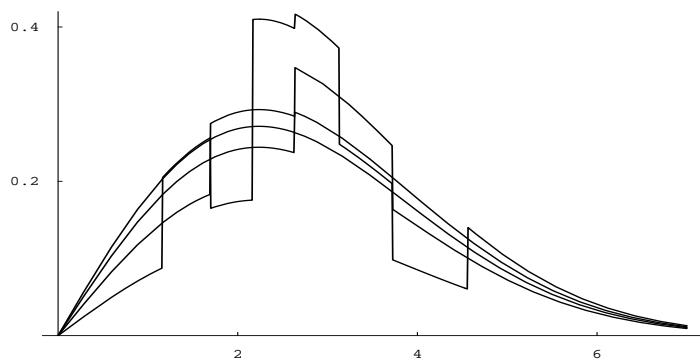
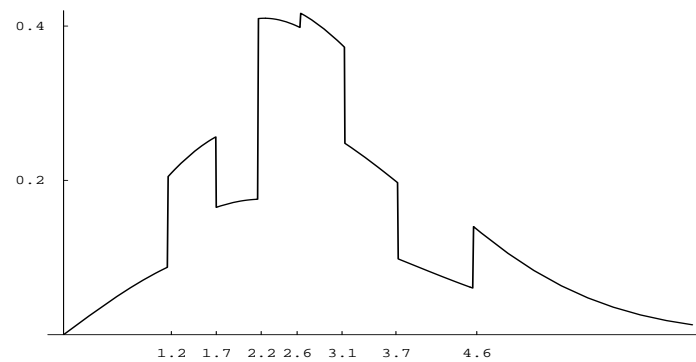
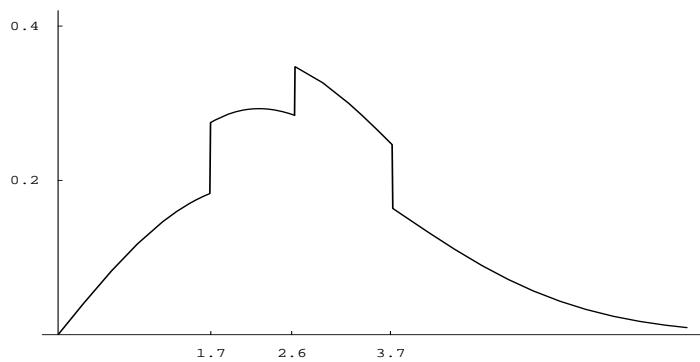
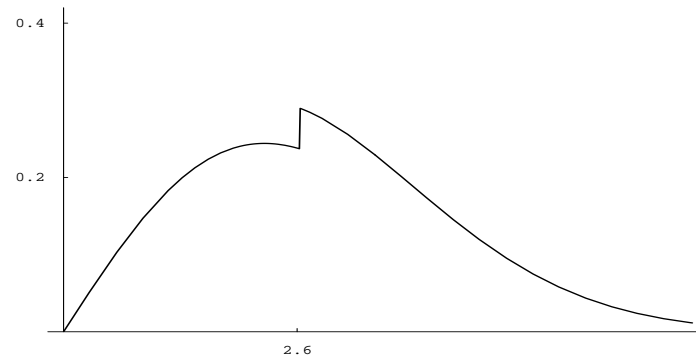
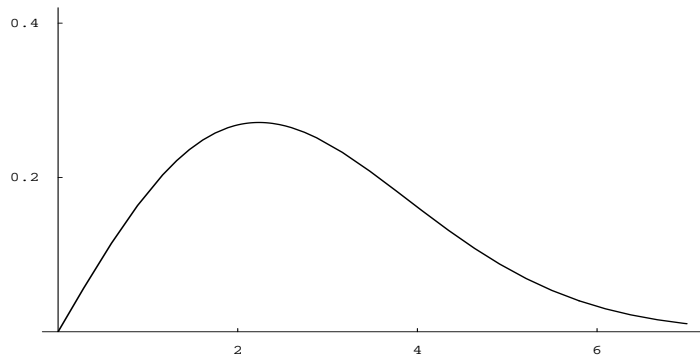


Figure 3:  $J = 0, 1, 2, 3$  for Weibull(2,10).

Prior on  $(Y_{\epsilon 0}, Y_{\epsilon 1})$

- Want  $E(Y_{\epsilon 0}) = 0.5$  to center  $G$  at  $G_{\theta}$ . Take

$$Y_{\epsilon 0} \sim \text{beta}(c\rho(j), c\rho(j)).$$

- $c$  and  $\rho(j)$  affect how quickly data “take over” parametric centering family  $\{G_{\theta} : \theta \in \Theta\}$ .
- $c$  is overall weight attached to  $G_{\theta}$ .  $\rho(j)$  affects how “clumped” data are.  $\rho(j) = j^2$  often used.
- Have  $E\{G(A)\} = G_{\theta}(A)$ .  $\text{var}\{G(A)\}$  depends on overall weight  $c$  and function  $\rho(\cdot)$ .

“Standard” parameterization where  $\rho(j) = j^2$ :

$\mathbb{R}$							
$B_0$					$B_1$		
$(Y_0, Y_1) \sim \text{Dir}(c, c)$							
$B_{00}$		$B_{01}$		$B_{10}$		$B_{11}$	
$(Y_{00}, Y_{01}) \sim \text{Dir}(4c, 4c)$				$(Y_{10}, Y_{11}) \sim \text{Dir}(4c, 4c)$			
$B_{000}$	$B_{001}$	$B_{010}$	$B_{011}$	$B_{100}$	$B_{101}$	$B_{110}$	$B_{111}$
$(Y_{000}, Y_{001}) \sim$		$(Y_{010}, Y_{011}) \sim$		$(Y_{100}, Y_{101}) \sim$		$(Y_{110}, Y_{111}) \sim$	
$\text{Dir}(9c, 9c)$		$\text{Dir}(9c, 9c)$		$\text{Dir}(9c, 9c)$		$\text{Dir}(9c, 9c)$	

$$\Pi_1 = \{B_0, B_1\}, \quad \mathcal{Y}_1 = \{Y_0, Y_1\}.$$

$$\Pi_2 = \{B_{00}, B_{01}, B_{10}, B_{11}\}, \quad \mathcal{Y}_2 = \{Y_{00}, Y_{01}, Y_{10}, Y_{11}\}.$$

$$\Pi_3 = \{B_{000}, B_{001}, B_{010}, B_{011}, B_{100}, B_{101}, B_{110}, B_{111}\}$$

$$\mathcal{Y}_3 = \{Y_{000}, Y_{001}, Y_{010}, Y_{011}, Y_{100}, Y_{101}, Y_{110}, Y_{111}\}$$

---

Adds 7 free parameters  $\mathcal{Y} = \{Y_0, Y_{00}, Y_{10}, Y_{000}, Y_{010}, Y_{100}, Y_{110}\}$ .

## Mixtures of Polya trees:

- MPT considered by Lavine (1992), Berger and Guglielmi (2001), Hanson and Johnson (2002), Hanson (2006), etc.
- Further taking  $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$  induces MPT; somewhat smooths 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.
- Next slide shows highly non-normal data and MPT estimate.  
 $BF \approx 10^{10}$  in favor of MPT vs.  $N(\mu, \sigma^2)$ .

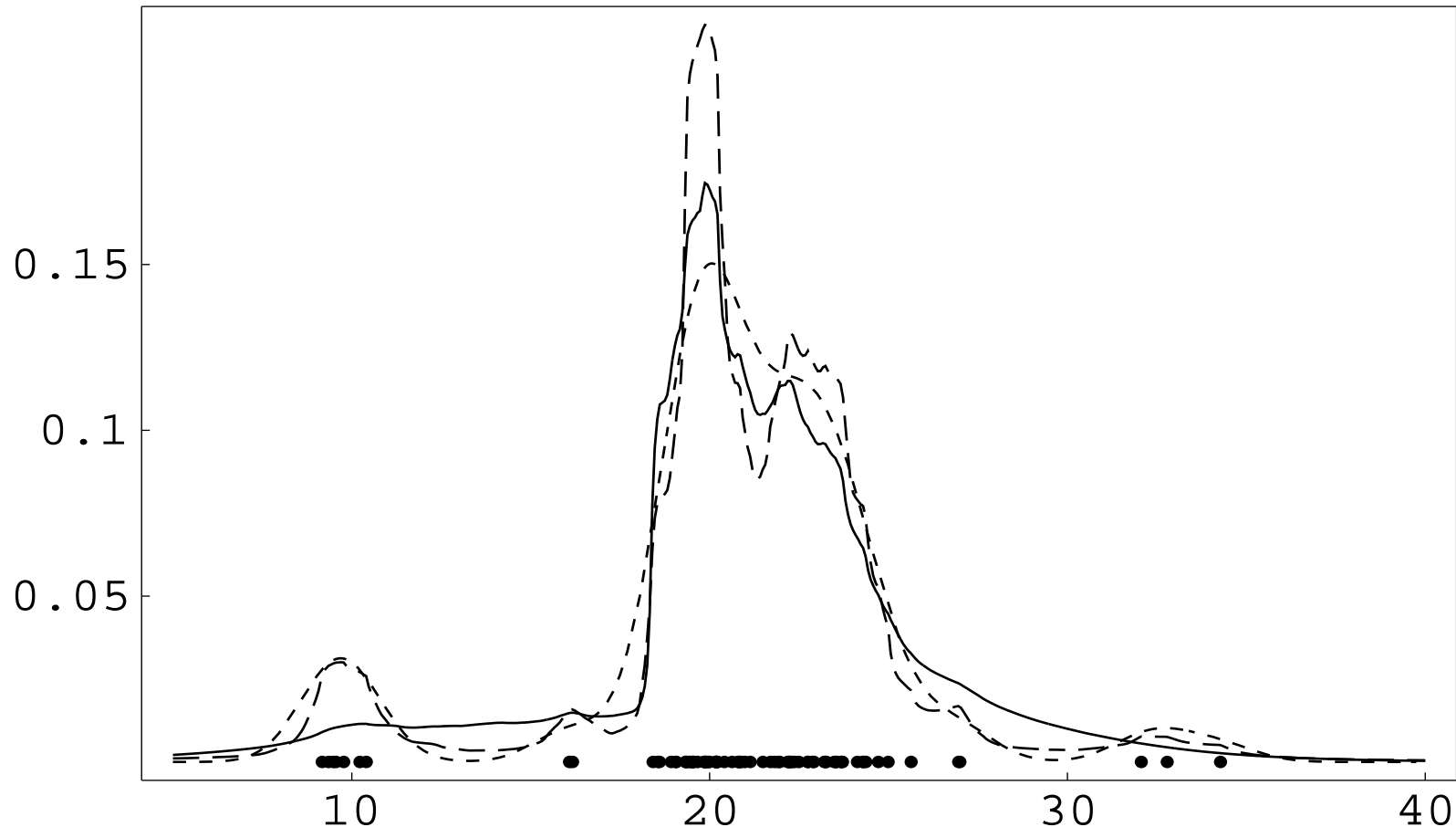


Figure 4: Galaxy data:  $x_1, \dots, x_{82} | G \sim G, G \sim PT_5(c, \rho, N(\mu, \sigma^2))$ . Solid is  $c \sim \Gamma(5, 1)$ , long-dashed is  $c = 0.1$ , short-dashed is Gaussian kernel-smoothed,  $h = 1$ .

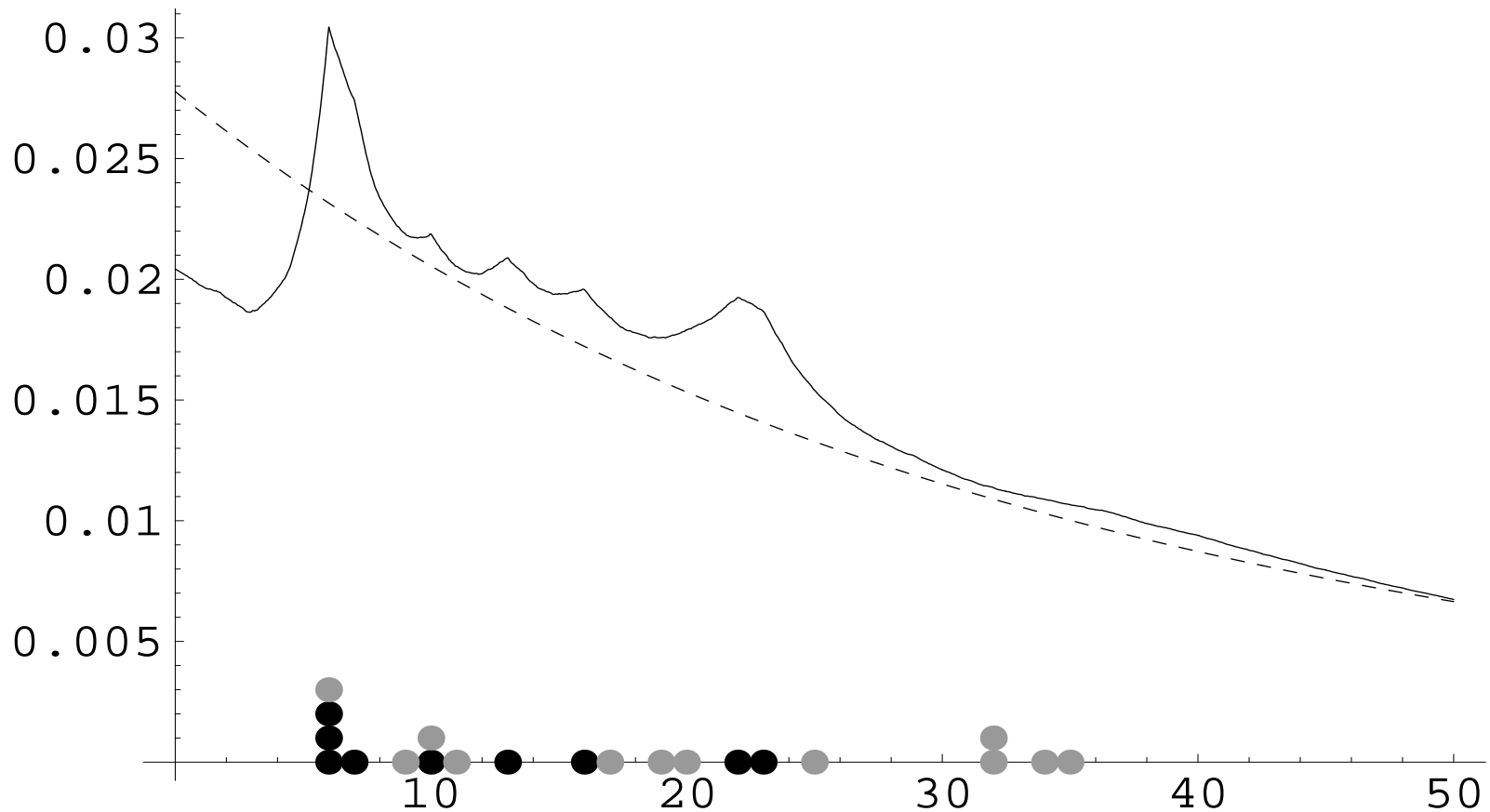


Figure 5:  $t_i$  = time to leukemia relapse; children treated w/ 6-mercaptopurine.  $t_1, \dots, t_{21} | G \stackrel{iid}{\sim} G$ ,  $G \sim \int PT_5(1, \rho, \exp(\theta)) p(\theta) d\theta$  where  $p(\theta) \propto 1$ . Density estimates for  $c = 1$  and  $c = \infty$ . Black uncensored, grey censored.  $PBF = 0.83$  for MPT vs.  $\exp(\theta)$ ,  $BF = 0.61$ .

Example 1:

Polya tree mixture model with  
application to meta analysis

## Meta analysis:

- Combine summaries across multiple studies designed to address the same scientific question.
- Observed study-specific effects  $\mathbf{y} = (y_1, \dots, y_n)$ ; corresponding 'true' effects  $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$ .
- e.g.  $\mathbf{y}$  vector of observed log odds ratios w/  $\boldsymbol{\theta}$  corresponding latent population LORs.
- $y_i$  approximately normal w/ mean  $\theta_i$  via asymptotics.
- Common to use  $N(\mu, \tau^2)$  for latent  $\theta_i$ .

## Normal-normal model

- For  $i = 1, \dots, n$ ,

$$\begin{aligned} y_i | \theta_i &\stackrel{ind.}{\sim} N(\theta_i, \sigma_i^2) \\ \theta_i | \mu, \tau &\stackrel{iid}{\sim} N(\mu, \tau^2) \end{aligned}$$

- $\sigma_i^2$  typically fixed at estimated variance of  $y_i$ .
- Normal-normal model attractive in part because  $\mu$  represents single encompassing effect measure.
- Primary statistical aim: make inferences about  $\mu$ , e.g.  $H_0 : \mu > 0$ .
- Study-specific covariates easily added.

## Generalize parametric normal-normal w/ MPT

- Main goal: broaden class of random effects distributions to include the normal but allow for flexibility. Retain simplicity of normal-normal model in terms of a unifying effect measure when appropriate
- Get direct, robust inference for functionals other than measures of center, which is particularly important when there is evidence of systematic heterogeneity across studies.
- Address computational aspects of model fitting, model selection, and nonparametric inferences for PTMs.

## Nonparametric meta-analysis: PTM

- Generalize second level of normal-normal using

$$\theta_i | G \stackrel{iid}{\sim} G$$

where  $G$  is assigned a Polya tree prior centered on the normal family.

$$y_i | \theta_i \stackrel{ind.}{\sim} N(\theta_i, \sigma_i^2)$$

$$\theta_i | G \stackrel{iid}{\sim} G$$

$$G | \mu, \tau, \nu, c \sim PT_J(c, \rho_\nu, N(\mu, \tau^2))$$

$$(\mu, \tau^2, \nu) \sim p(\mu, \tau^2, \nu)$$

where  $\sigma_1, \dots, \sigma_n$  are fixed constants.

- $\mu$  is still the median; MPT constrained so that  $G(\mu) = 0.5$ .

## Nonparametric meta-analysis: PTM

- Prior is

$$\mu|\tau^2 \sim N(a_\mu, b_\mu\tau^2), \quad \tau^{-2} \sim \Gamma(a_\tau, b_\tau), \quad \nu \sim N(a_\nu, b_\nu).$$

- Weight  $c$  set to small constant, e.g.  $c = 1$ . Too small gives prior mass on approximately discrete distributions  $\Rightarrow$  MCMC mixing suffers to point of intractability.
- Posterior inferences for median  $\mu$  and  $G$  obtained from Gibbs sampling.

## Nonparametric meta-analysis: DPM

- PTM model provides alternative to Dirichlet process mixture (DPM) model.
- DPM adds flexibility by instead considering  $G|\mu, \tau \sim DP(\alpha, N(\mu, \tau^2))$ ; i.e. a DP with fixed weight  $\alpha$  and prior on  $(\mu, \tau^2)$ .
- Use of a DP random effects distribution eliminates the straightforward interpretation of  $\mu$  since it no longer represents an encompassing effect measure.
- To fix this, Burr and Doss (2005, JASA) consider a conditional DPM in which  $G$  has median  $\mu$ .

## Polya tree mixture models

- For Polya tree priors, typically  $\rho(j) = j^2$  giving absolutely continuous  $G$  with probability one in fully specified tree.
- Unlike previous approaches, we consider

$$\rho_\nu(j) = 2^{-\nu j}.$$

Allows for spikes and clumping of probability mass; e.g. can approximate discrete  $G$ .

- e.g.  $\nu = 1$  gives approximate DP prior;  $\nu < 0$  gives densities.

## Model comparison

- Empirical Bayes approach coupled w/ sensitivity analysis.  
 $c = 1, 10$  &  $\nu$  is fixed at a posterior estimate.
- For fixed  $(c, \nu)$ , a test of whether or not the underlying Gaussian model is appropriate is carried out using the Savage-Dickey ratio.
- The normal model obtains when PT conditional probabilities are all 0.5.

## Model comparison

- Bayes factors.
- Also log-pseudo marginal likelihood (LPML): measure of how well supported each observation is by the remaining data and model, aggregated over all  $n$  observations

$$\text{LPML} = \sum_{i=1}^n \log p(y_i | \mathbf{y}_{-i})$$

- Compare models using the pseudo Bayes factor, which roughly indicates which model is superior at predicting the observed data:  $\text{PBF}_{12} = \exp(\text{LPML}_1 - \text{LPML}_2)$

## Simulated Data

- Data: Sample means simulated from  $n = 100$  studies each w/ 100 individuals.
- $\theta_1, \dots, \theta_{100} \sim$  skewed bimodal distribution  $\Rightarrow$  get inferences for  $G$  since this suggests systematic heterogeneity across location.
- Model:  $\bar{y}_i | \theta_i \sim N(\theta_i, \sigma^2/100)$ ,  $\theta_i \sim G$  where  $G \sim PT(c, \rho_\nu, N(\mu, \tau^2))$
- Prior:  $\mu | \tau^2 \sim N(0, 1000\tau^2)$  and  $\tau^{-2} \sim \Gamma(0.01, 0.01)$ . Set  $c = 1$  and used  $\nu = -1$  for calculating BF, and  $\nu \sim N(0, 4)$  for inference about functionals of  $G$
- Sampled 100 simulated data sets of size 100.
- Bayes factors ranged from  $10^5$  to  $10^{20}$  in favor of the PTM.

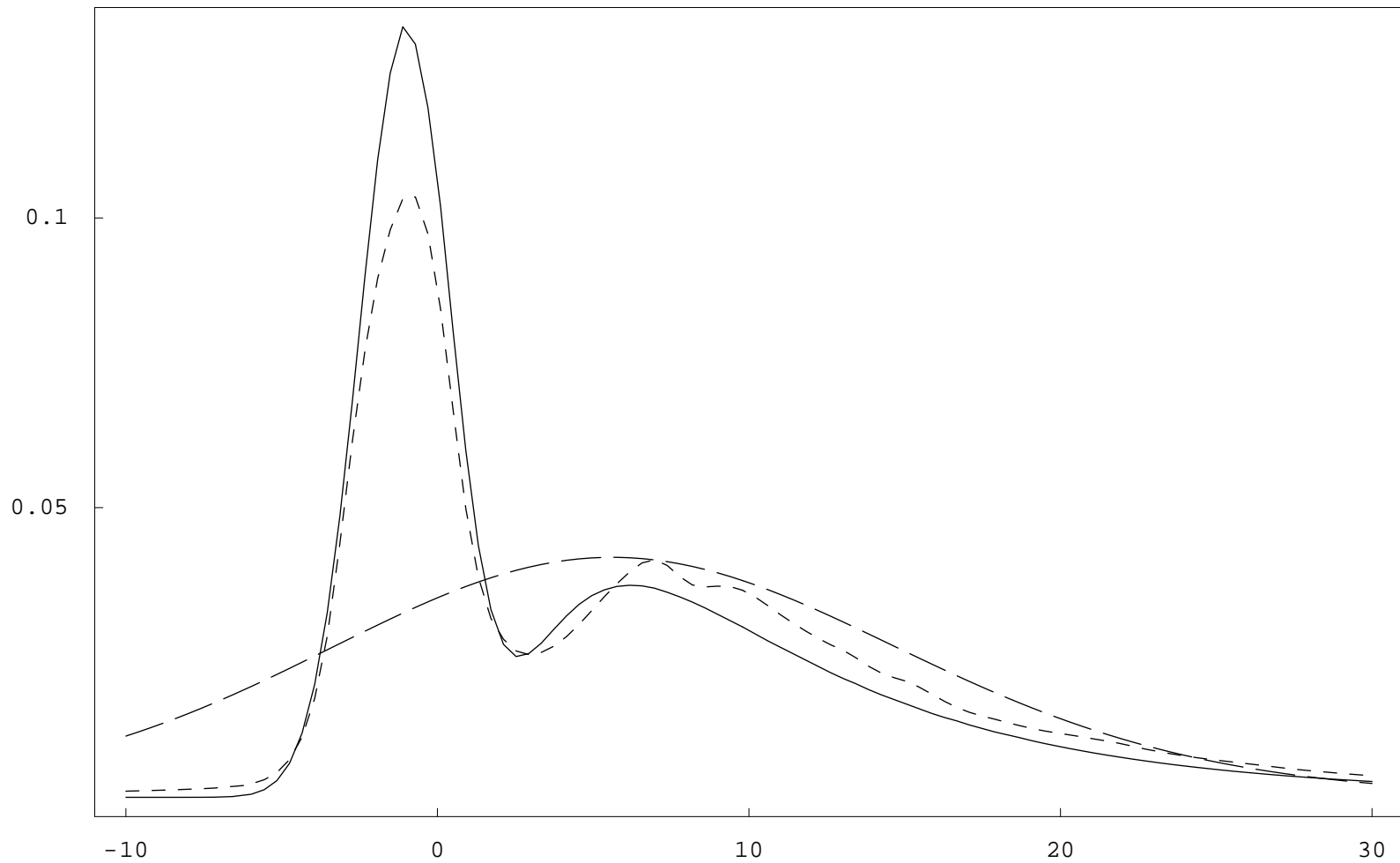


Figure 6: Normal-normal fit & PTM refinement. Dashed = average posterior across 100 data sets. True effects dist'n = solid line.

## Alcohol and breast cancer

- Meta-analysis of 39 studies; mix of retrospective and prospective designs; conducted in several countries.
- Summary measure was the estimated change in log odds ratio (scaled as  $LOR \times 1000$ ) for a one gram increase in daily alcohol consumption.
- PTM analysis:  $J = 5$ ;  $\mu|\tau^2 \sim N(0, 1000\tau^2)$ ,  $\tau^{-2} \sim \Gamma(0.01, 0.01)$ , and  $\nu \sim N(0, 4)$ . With these priors, the posterior mode of  $\nu$  is essentially 0 so we also analyzed the data and computed PBFs with  $\nu$  fixed at 0.
- PTM analysis favored over the normal-normal model with PBF  $\approx 3000$ .
- A comparison of the conditional DPM ( $\alpha = 1$ ) versus PTM model yielded PBF  $\approx 220$  in favor of the PTM.

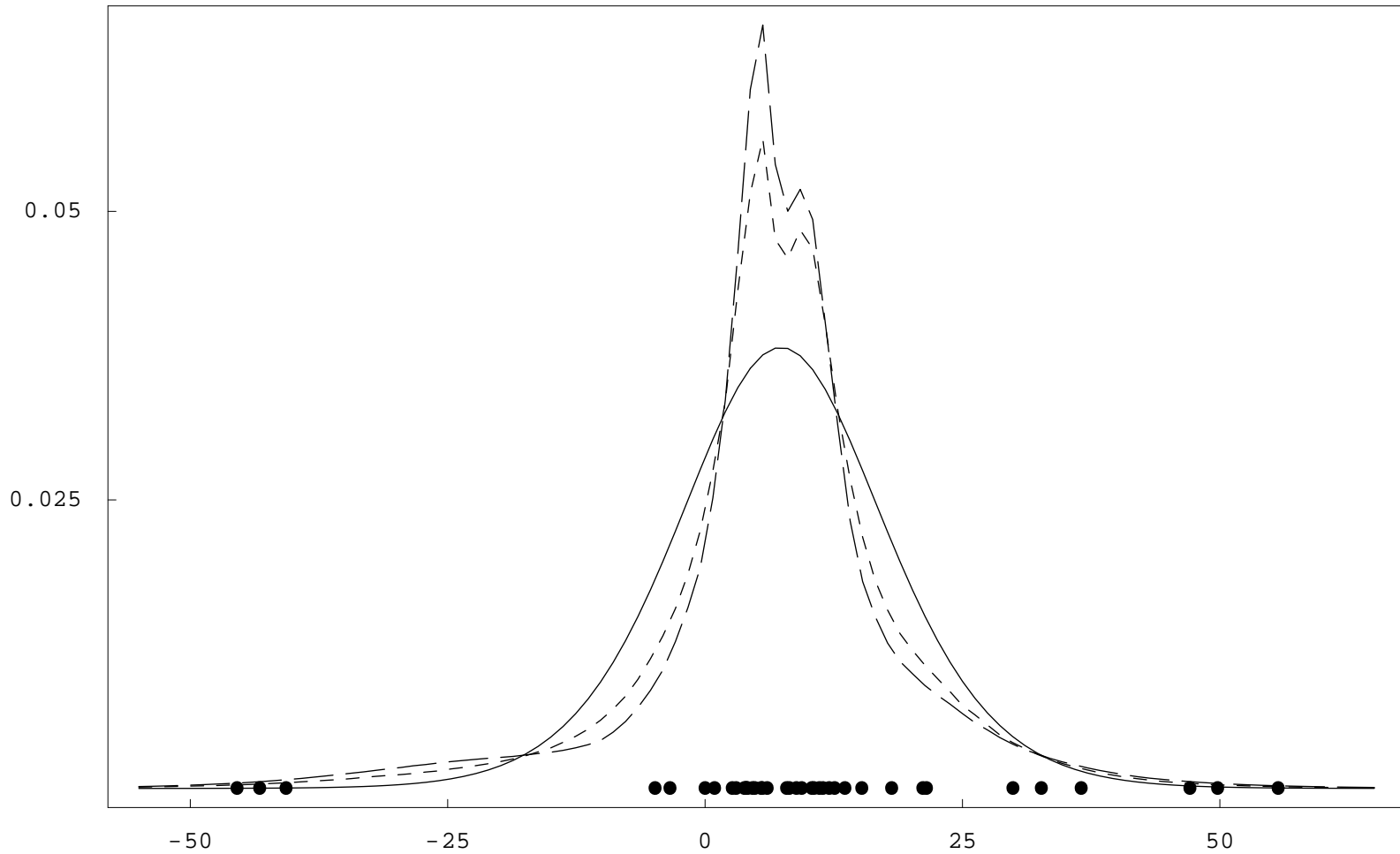


Figure 7: Data-driven flexibility of the PTM apparent, especially in the outer percentiles. Short-dashed  $\nu \sim N(0, 4)$ ; long-dashed  $\nu = 0$ .

Example 2:

Cox regression with nonparametric  
frailties fit in **DPpackage**

**Generalized linear mixed models:** generalizing Gaussian random effects distribution

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  $G$ .

## Frailty modeling in the proportional hazards model

Hazard function:

$$\lambda_0(t) = \lim_{\epsilon \rightarrow 0} \frac{P(t \leq T_0 < t + \epsilon | T_0 \geq t)}{\epsilon} = \frac{f_0(t)}{S_0(t)}.$$

Conditionally proportional hazards:

$$\lambda(t_{ij} | \mathbf{x}_{ij}) = \lambda_0(t) \exp(\mathbf{x}'_{ij} \boldsymbol{\beta} + \gamma_i).$$

- $i = 1, \dots, n$  groups
- $j = 1, \dots, n_i$  within group  $i$ .
- Data  $\{(\mathbf{x}_{ij}, t_{ij}, \delta_{ij})\}$ .

- Hazard function is piecewise constant on partition of  $\mathbb{R}^+$ .  
Partition comprised of  $K$  intervals.

$$\lambda_0(t) = \sum_{k=1}^K \lambda_k I\{a_{k-1} < t \leq a_k\}$$

where  $a_0 = 0$  and  $a_K = \infty$ .

- Implies Poisson likelihood

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\gamma}) = \prod_{i=1}^n \prod_{j=1}^{n_i} \left[ \prod_{k=1}^{K(t_{ij})} e^{-e^{\mathbf{x}'_{ij}\boldsymbol{\beta} + \gamma_i} \Delta_{ijk}} \right] \left[ e^{\mathbf{x}'_{ij}\boldsymbol{\beta} + \gamma_i} \lambda_{K(t_{ij})} \right]^{\delta_{ij}}$$

where  $\Delta_{ijk} = \min\{a_k, t_{ij}\} - a_{k-1}$  and  $K(t) = \max\{k : a_k \leq t\}$ .

- Take  $K = 10$  from %iles of data.

- Data on  $n = 38$  kidney patients; each has  $n_i = 2$  infection times, several right censored.
- Only covariate of interest: gender.
- Analyzed in Walker and Mallick (1997) with piecewise exponential model and frailties following  $PT_8(0.1, \rho, N(0, 10^2))$ .
- Poisson likelihood in form of GLMM. Fit in DPpackage:
 

```
fit2=PTglmm.default(fixed = y ~ g + h, random = ~1 | id,
family = poisson(log), offset = log(off), prior = prior2,
mcmc = mcmc2, state = state, status = FALSE)
```
- Assume

$$\gamma_1, \dots, \gamma_n | G \stackrel{iid}{\sim} G, \quad G \sim PT(c, \rho, N(0, \sigma^2)), \quad \sigma^{-2} \sim \Gamma(a_\sigma, b_\sigma)$$

$$\log \boldsymbol{\lambda} \sim N_K(\mathbf{m}_0, \mathbf{V}_0), \quad \boldsymbol{\beta} \sim N_p(\mathbf{b}_0, \mathbf{B}_0), \quad c \sim \Gamma(a_c, b_c).$$

Bayesian semiparametric generalized linear mixed effect model

Model's performance:

Dbar	Dhat	pD	DIC	LPML
374.45	351.09	23.37	397.82	-201.66

Regression coefficients:

	Mean	Median	Std. Dev.	Naive Std.Error	95%CI-Low	95%CI-Upp
(Intercept)	-0.013506	-0.006234	0.142077	0.004493	-0.320878	0.277733
g	-1.435111	-1.416729	0.482620	0.015262	-2.402870	-0.549498
h1	-4.299368	-4.259838	0.571100	0.018060	-5.509669	-3.298870
h2	-3.740145	-3.702957	0.576537	0.018232	-4.907587	-2.682208
h3	-3.907660	-3.876771	0.570570	0.018043	-5.100021	-2.861654
h4	-2.948614	-2.912574	0.573208	0.018126	-4.179309	-1.892098
h5	-3.087463	-3.046317	0.546165	0.017271	-4.138270	-2.070438
h6	-3.748874	-3.747799	0.581659	0.018394	-4.987058	-2.621747
h7	-4.597002	-4.558640	0.636655	0.020133	-5.867402	-3.480104
h8	-3.324208	-3.332950	0.602217	0.019044	-4.503470	-2.208580
h9	-3.218744	-3.234147	0.666629	0.021081	-4.549745	-1.911680
h10	-3.515208	-3.547174	0.687001	0.021725	-4.857640	-2.108612

Baseline distribution:

	Mean	Median	Std. Dev.	Naive Std.Error	95%CI-Low	95%CI-Upp
mu-(Intercept)	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
sigma-(Intercept)	0.65845	0.52536	0.53939	0.01706	0.15881	1.91989

Precision parameter:

	Mean	Median	Std. Dev.	Naive Std.Error	95%CI-Low	95%CI-Upp
alpha	0.3957673	0.3974533	0.0045467	0.0001438	0.3843464	0.4009077

Acceptance Rate for Metropolis Steps = 0.1933333 0.6004101 0 0.1512987 0.994026

Density of (Intercept)

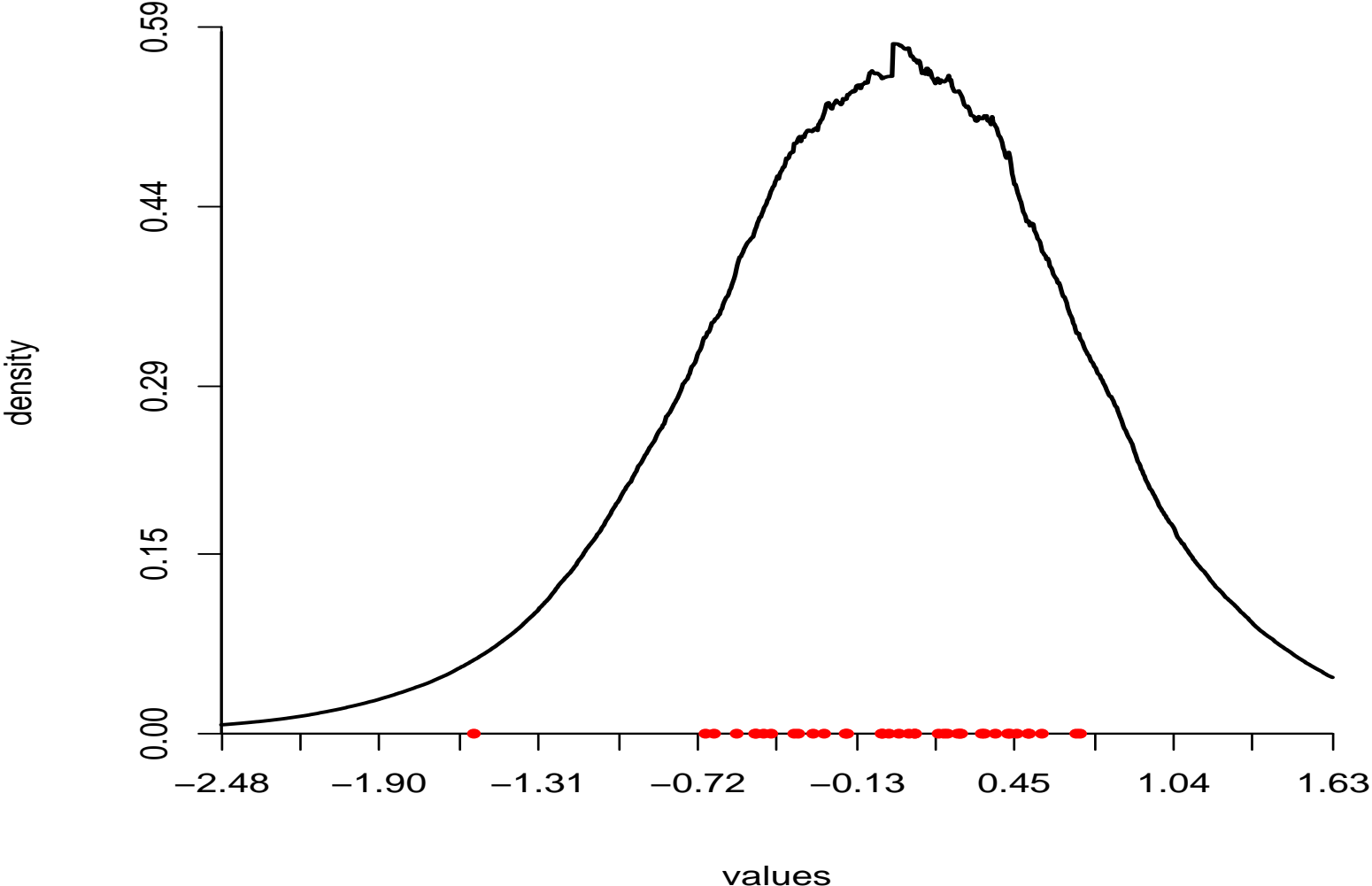


Figure 8: Estimated frailty distribution.

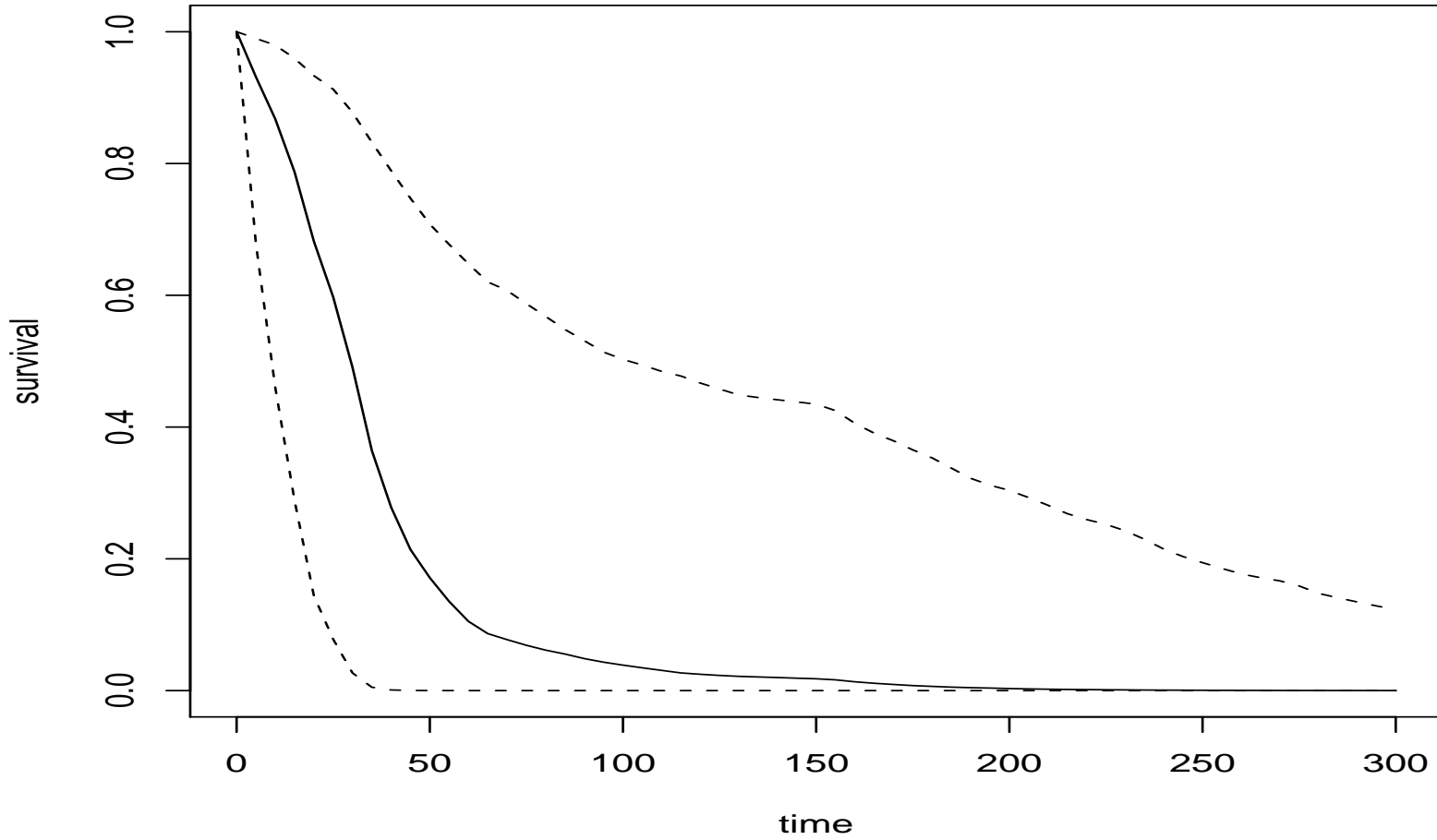


Figure 9: Predictive time to infection: males.

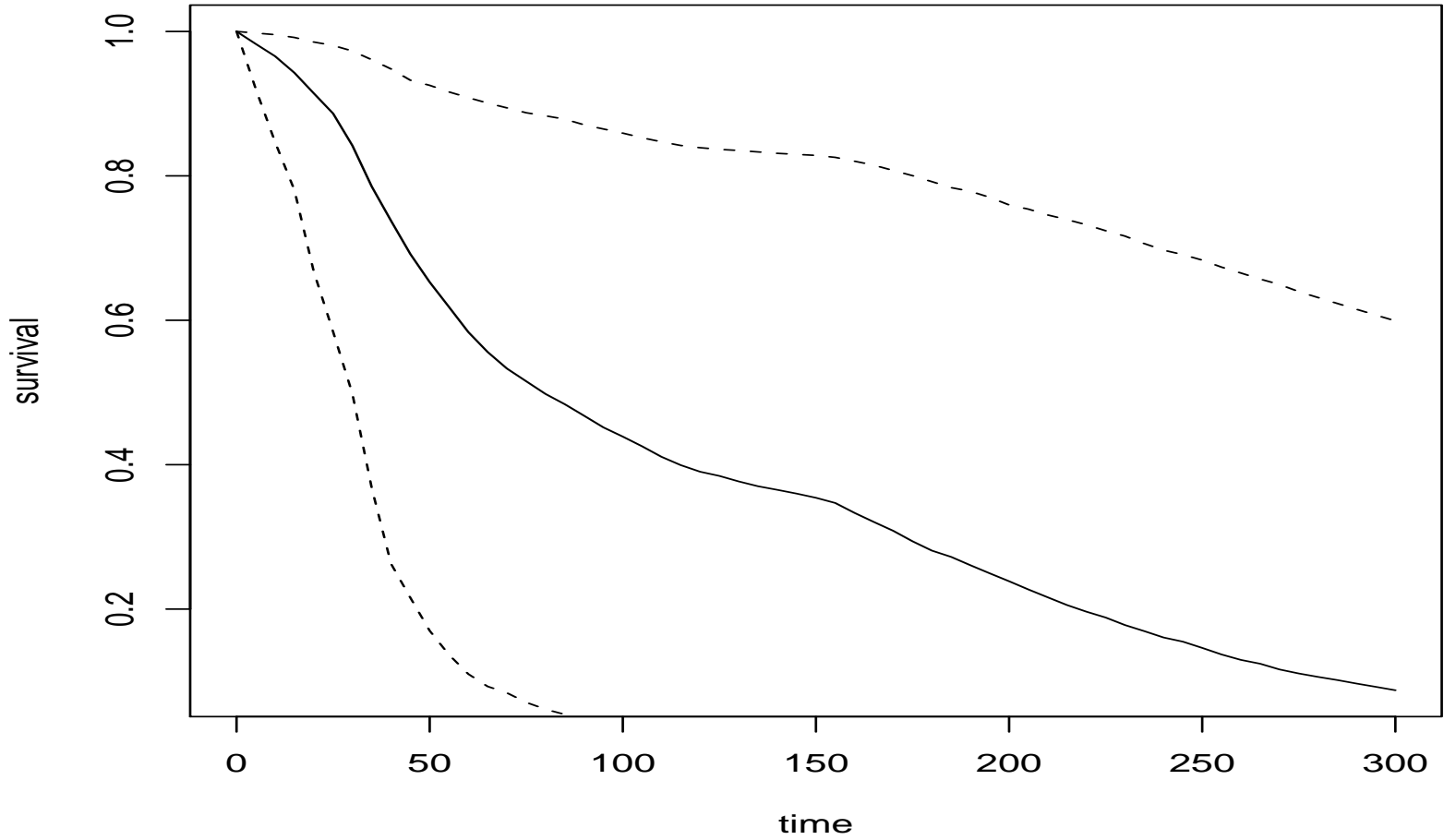


Figure 10: Predictive time to infection: females.

- We obtain  $\hat{\beta}_g = -1.4$ . McGilchrist and Aisbett (1991) get  $\hat{\beta}_g = -1.8$  w/ other covariates included. Aslanidou et al. (1995) and Walker and Mallick (1997) get  $\hat{\beta}_g = -1.0$ .
- DIC = 398 from either MPT or normal model. Normal model does about the same.
- MPT generalization doesn't do harm if not needed.

Example 3:

Partially exchangeable random effects with application to Ache monkey hunting

- Paraguayan Ache tribe part-time hunter-gatherers; in contact with “modern civilization” only since the mid-1970’s.
- McMillan (2001) spent year living w/ Ache collecting data.
- Part of Ache life spent on extended forest treks – only eat food they gather or hunt. Capuchin monkeys shot out of trees.
- Hunters split into two groups, one chasing a troop of monkeys towards the other group who shoot at them with bows and arrows.
- Dangerous because arrows fired straight up fall back out of the trees. Hunting prowess contributes to group status.
- *Questions:* how does age affect ability to hunt monkeys? How heterogenous is hunting ability?

- $Y_{ij}$  = monkeys killed by hunter  $i$  on  $j$ th trek,  
 $i = 1, \dots, 47, j = 1, \dots, N_i$ .
- $M_{ij}$  = length in days trek  $ij$ .
- Model:

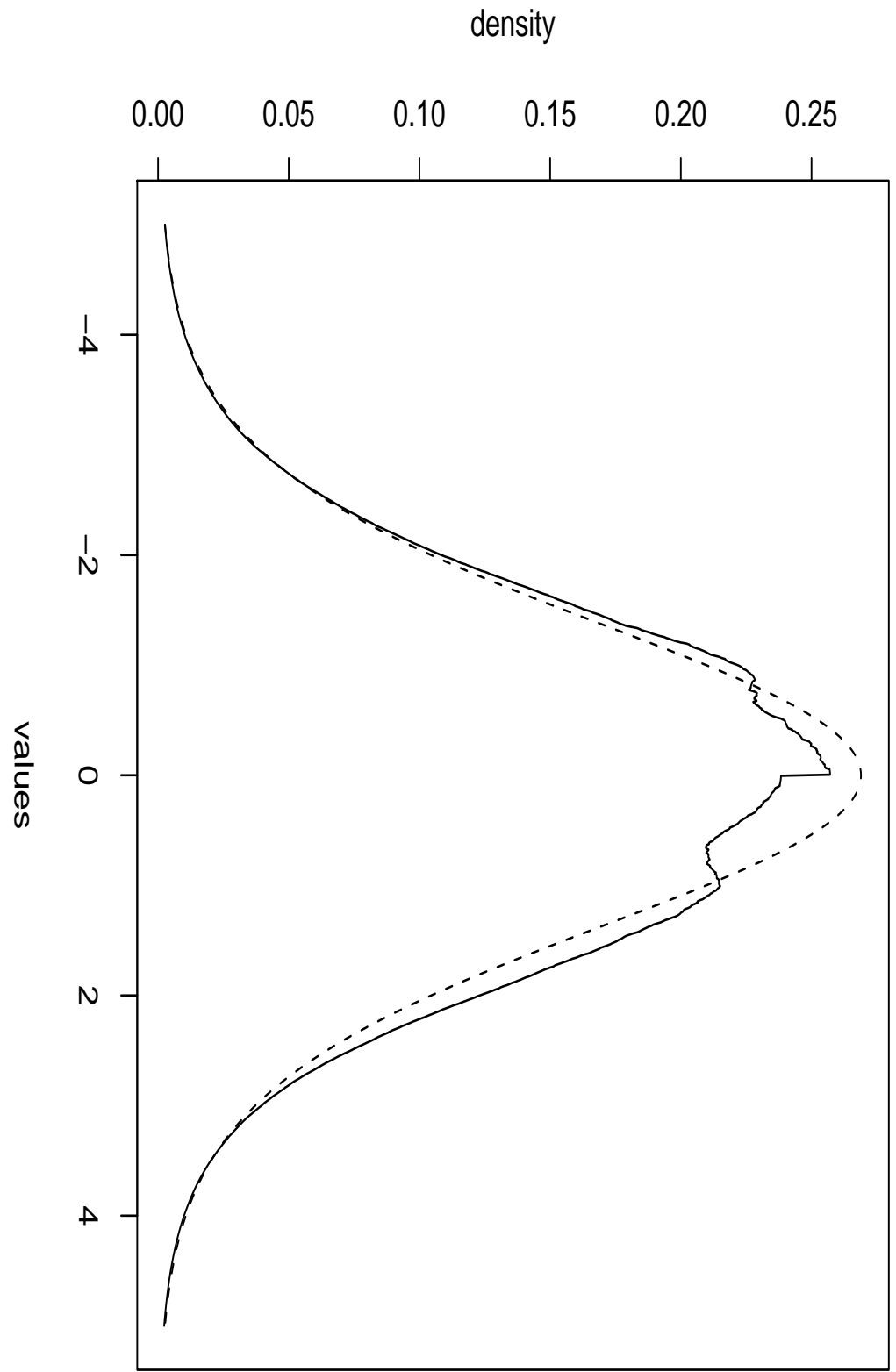
$$Y_{ij} | \lambda_i \sim \text{Pois}(\lambda_i M_{ij}),$$

where

$$\log \lambda_i = \beta_0 + \beta_1(a_i - 45) + \beta_2(a_i - 45)^2 + \gamma_i.$$

- Fixed effects  $(\beta_0, \beta_1, \beta_2)$  capture overall population trend in hunting ability due to age  $a_i$ .
- Random effects  $\gamma_i$  are surrogate for “innate” hunting ability.
- Random effects commonly assumed exchangeable normal

$$\gamma_1, \dots, \gamma_{47} | \sigma \sim N(0, \sigma^2).$$



- McMillan, anthropologist who spend a year w/ Ache surprised.
- Expected “clumps” of good and bad hunters based on experience.
- We see essentially Gaussian distribution of hunting ability.
- Problem: confounding between frailty distribution and hunter’s age. Older hunters spent more time in the forest *when they were young*; younger Ache spend more time farming and playing soccer.
- Age represents warped timescale of *actual time spend hunting*, which would be better predictor. Unfortunately cannot easily measure this.
- New model  $\gamma_i | \mathbf{x}_i \sim G_{\mathbf{x}_i}$  where  $G_{\mathbf{x}_i}$  has  $Y_\epsilon = Y_\epsilon(a_i)$ .

## Dependent Polya tree

- Basic idea is simple: replace each conditional probability w/ a logistic regression:

$$Y_{\epsilon 0}(\mathbf{x}, \boldsymbol{\beta}_{\epsilon 0}) = \frac{\exp(\mathbf{x}'\boldsymbol{\beta}_{\epsilon 0})}{1 + \exp(\mathbf{x}'\boldsymbol{\beta}_{\epsilon 0})}, \quad Y_{\epsilon 1}(\mathbf{x}, \boldsymbol{\beta}_{\epsilon 0}) = \frac{1}{1 + \exp(\mathbf{x}'\boldsymbol{\beta}_{\epsilon 0})}.$$

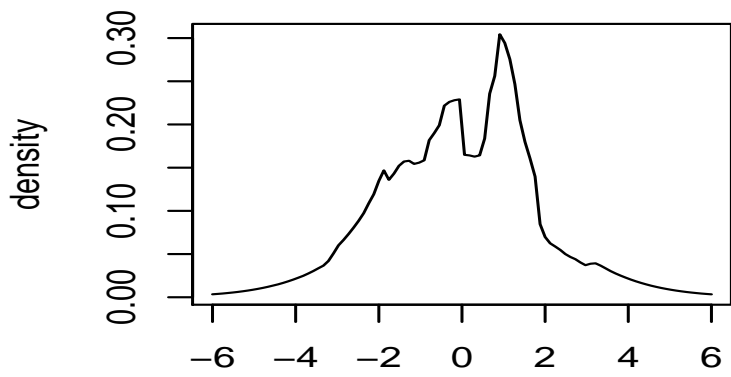
- $\epsilon_{\theta}(\gamma, j)$  is set in  $\Pi_j^{\theta}$  that  $\gamma$  is in.

$$g(\gamma_i) = 2^J \phi_{\theta}(\gamma_i) \prod_{i=1}^J Y_{\epsilon_{\theta}(\gamma_i, j)}.$$

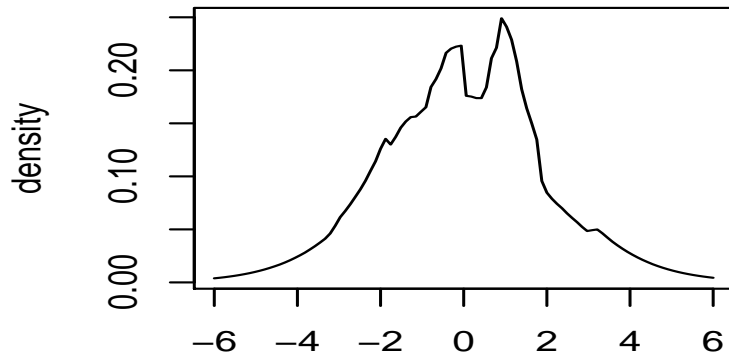
- Density of random effects is

$$p(\gamma_1, \dots, \gamma_n | \theta, \boldsymbol{\beta}) = \left[ \prod_{i=1}^n \phi_{\theta}(\gamma_i) \right] \left[ \prod_{\epsilon \in E^{J-1}} \prod_{i=1}^n \frac{\exp(\mathbf{x}'_i \boldsymbol{\beta}_{\epsilon 0})^{I\{\gamma_i \in B_{\theta}(\epsilon 0)\}}}{[1 + \exp(\mathbf{x}'_i \boldsymbol{\beta}_{\epsilon 0})]^{I\{\gamma_i \in B_{\theta}(\epsilon)\}}} \right].$$

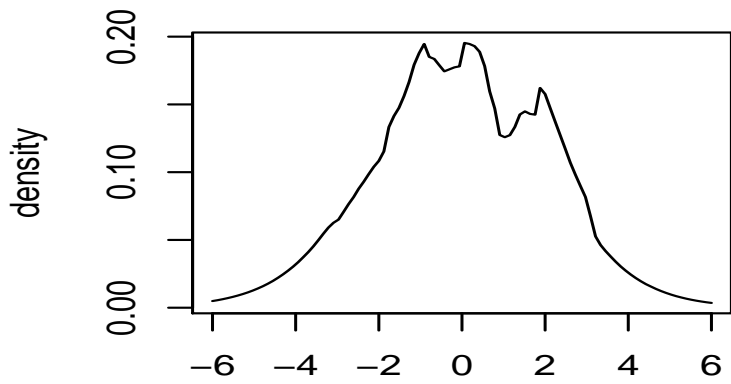
- Product of  $2^J - 2$  logistic regression kernels, one for each  $\epsilon 0$ , times parametric likelihood for  $\theta$ .



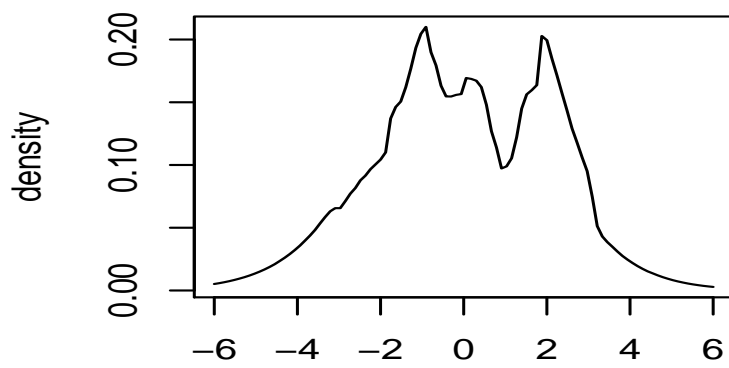
frailty density: age=17



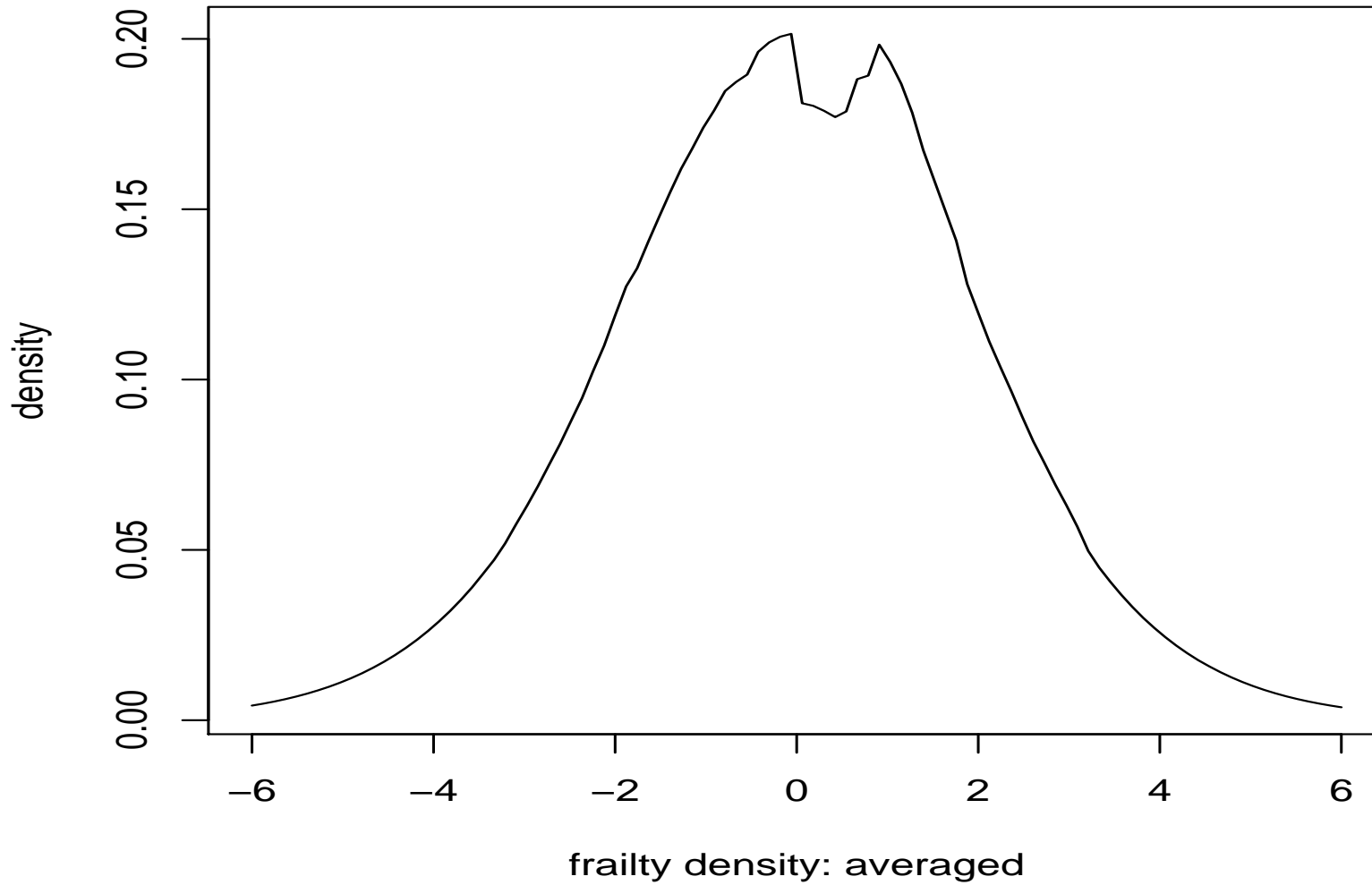
frailty density: age=36



frailty density: age=58



frailty density: age=73



- Inference surprisingly easy to get: uses iteratively re-weighted least squares M-H proposal (Gamerman, 1997).
- Evidence of confounding. Density ‘evolves’ with age. Shows more heterogeneity and clump of ‘better’ hunters for older ages.
- Densities averaged over age looks almost identical to exchangeable PT density!
- Dependent model provides best prediction according to LPML.  $PBF \approx 50$  for DTFP vs. Gaussian.

Model	LPML	$\beta_0$	$\beta_1$	$\beta_2$
$N(0, \sigma^2)$	-69	-2.54	0.041	-0.0061
$PT(c, \rho, N(0, \sigma^2))$	-70	-2.62	0.042	-0.0057
$DTFP(c, \rho, N(0, \sigma^2))$	-65	-3.00	0.058	-0.0073

- Will eventually be in DPpackage

```
fit=DTFglm(fixed = monkeys ~ ages1 + ages2, random = ~1 | hunter,  
  modeltf = ~ages1, family = poisson(log), offset = log(days),  
  prior = prior, mcmc = mcmc, state = NULL, status = TRUE,  
  prediction = prediction)
```

- Other dependent Polya tree models use logit-transformed Gaussian process for conditional probabilities and CAR prior giving longitudinal and spatial dependence, respectively.
- Applications to modeling breast cancer survival from SEER data: two Minnesota Biostat tech reports.

Example 4:

A bit on multivariate Polya trees

- 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  $N_d(\mathbf{0}, \mathbf{I})$ .
- Considering  $\mathbf{U} = \mathbf{M}\mathbf{\Lambda}^{1/2}\mathbf{O}$  where  $\mathbf{O} \sim \text{Haar}(d)$  vastly smooths inference (Jara et al., 2009).
- 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  $10^{200}$ .

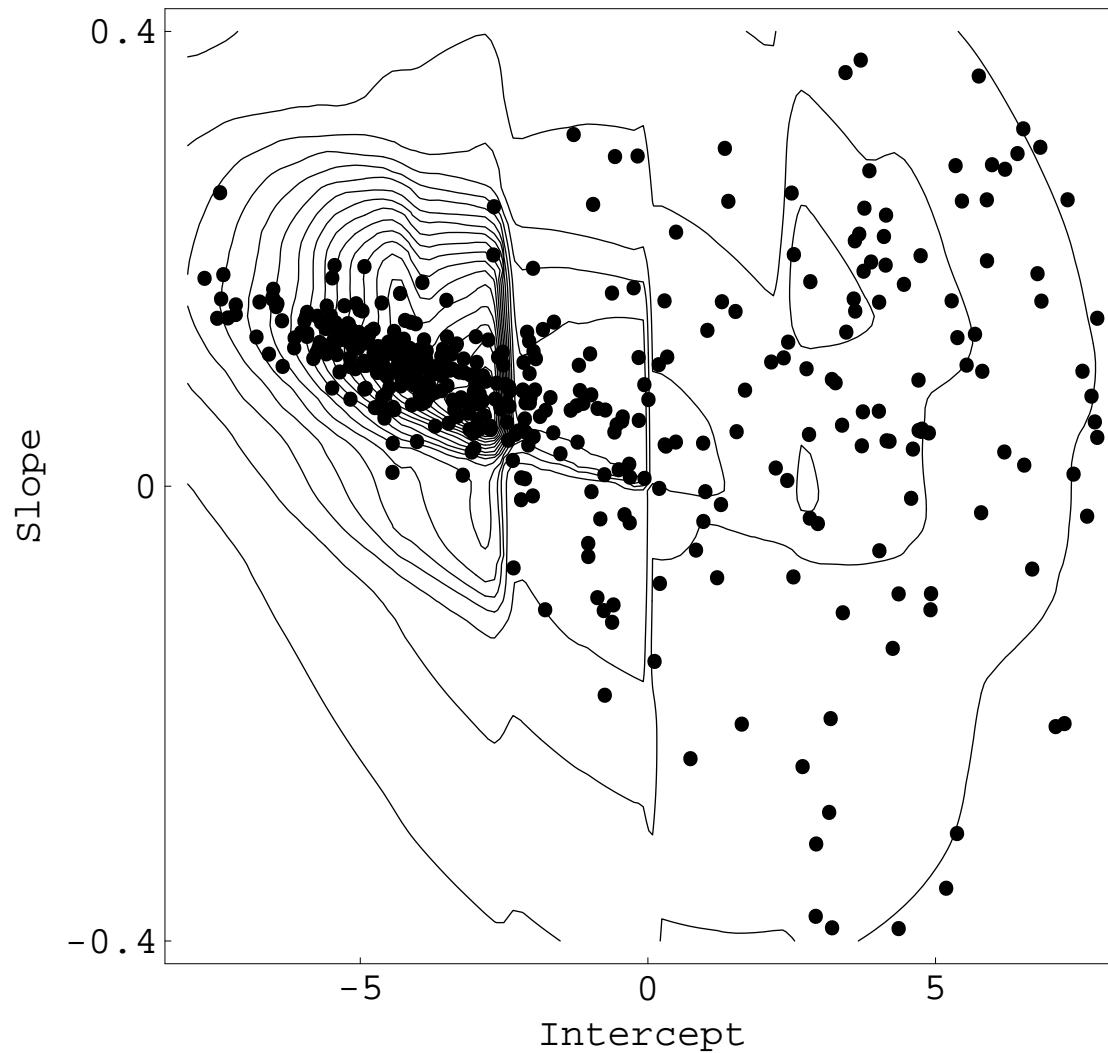


Figure 11: Predictive density level curves of  $G$  with  $E(\mathbf{b}_i | \mathbf{y}_1, \dots, \mathbf{y}_{467})$ .

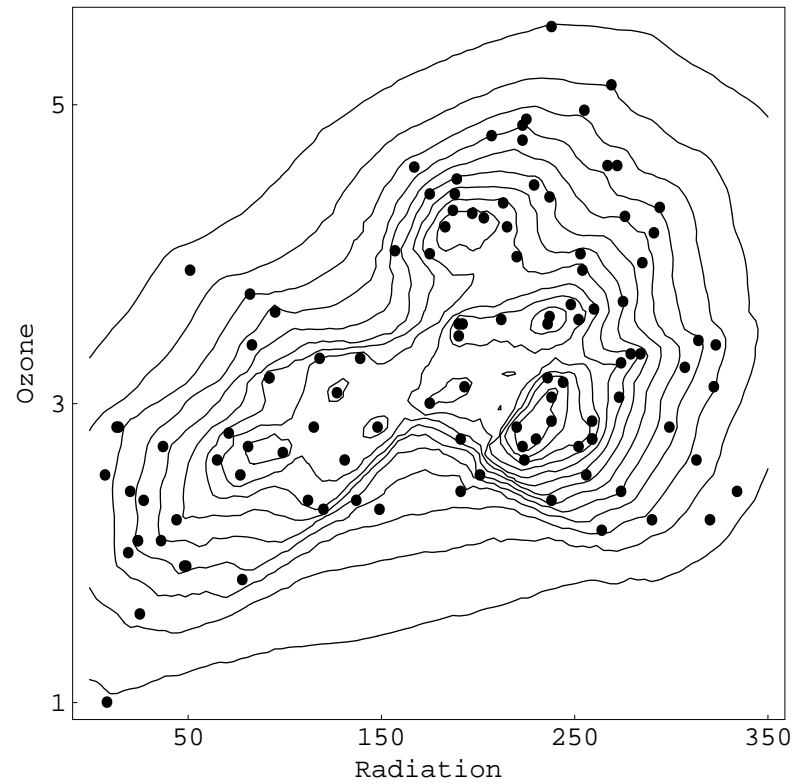
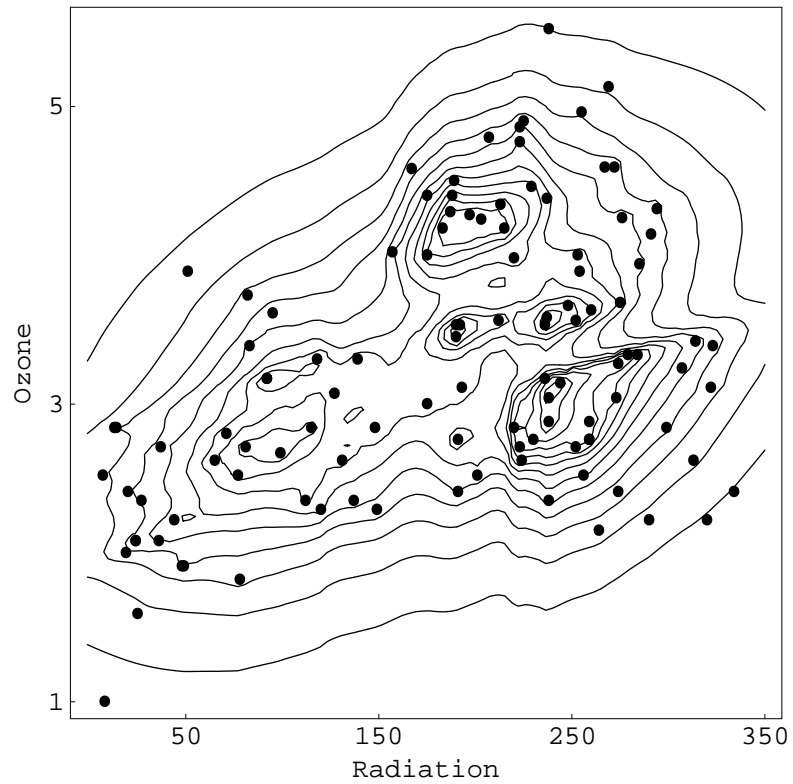


Figure 12: Last plot! Environmental data:  $\mathbf{U} = \mathbf{M}\mathbf{\Lambda}^{1/2}\mathbf{M}'$  (left) and  $\mathbf{U} = \mathbf{M}\mathbf{\Lambda}^{1/2}\mathbf{O}$ ,  $\mathbf{O} \sim \text{Haar}(d)$  (right).

## Comments...

- MPT have not been tapped to full potential.
- 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 BF or PBF.
- Current research: fast MMPT approximations to posterior densities, dependent processes, multivariate survival analysis, ordinal regression.
- Thanks!