AMS 241: Bayesian Nonparametric Methods
Notes 2 – Dirichlet process mixture models

Instructor: Athanasios Kottas

Baskin School of Engineering
University of California, Santa Cruz
Spring 2018

Outline

1. Introduction and motivation
2. Dirichlet process mixtures: definitions and model properties
3. Posterior simulation methods
4. Applications

Motivating Dirichlet process mixtures

- Recall that the Dirichlet process (DP) is a conjugate prior for random distributions under i.i.d. sampling.
- However, posterior draws under a DP model correspond (almost surely) to discrete distributions. This is somewhat unsatisfactory if we are modeling continuous distributions.
- In the spirit of kernel density estimation, one solution is to use convolutions to smooth out posterior estimates.
- In a model-based context, this leads to DP mixture models, i.e., a mixture model where the mixing distribution is unknown and assigned a DP prior (recall that this is different from a mixture of DPs, in which the parameters of the DP are random).
- Strong connection with finite mixture models.
- More generally, we might be interested in using a DP as part of a hierarchical Bayesian model to place a prior on the unknown distribution of some of its parameters (e.g., random effects models). This leads to semiparametric Bayesian models.

Mixture distributions

- Mixture models arise naturally as flexible alternatives to standard parametric families.
- Continuous mixture models (e.g., t, Beta-binomial, and Poisson-gamma models) typically achieve increased heterogeneity but are still limited to unimodality and usually symmetry.
- Finite mixture distributions provide more flexible modeling, and are now relatively easy to implement, using simulation-based model fitting (e.g., Richardson and Green, 1997; Stephens, 2000; Jasra, Holmes and Stephens, 2005).
- Rather than handling the very large number of parameters of finite mixture models with a large number of mixture components, it may be easier to work with an infinite dimensional specification by assuming a random mixing distribution, which is not restricted to a specified parametric family.

Finite mixture models

- Recall the structure of a finite mixture model with K components, for example, a mixture of $K = 2$ Gaussian densities:

$$y_i \mid w, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \overset{ind.}{\sim} w\mathrm{N}(y_i \mid \mu_1, \sigma_1^2) + (1-w)\mathrm{N}(y_i \mid \mu_2, \sigma_2^2),$$

  that is, observation y_i arises from a $\mathrm{N}(\mu_1, \sigma_1^2)$ distribution with probability w or from a $\mathrm{N}(\mu_2, \sigma_2^2)$ distribution with probability $1 - w$ (independently for each $i = 1, \ldots, n$, given the parameters).

- In the Bayesian setting, we also set priors for the unknown parameters

$$(w, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \sim p(w, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2).$$

Finite mixture models

- The model can be rewritten in a few different ways. For example, we can introduce auxiliary random variables $L_1, \ldots, L_n$ such that $L_i = 1$ if y_i arises from the $\mathrm{N}(\mu_1, \sigma_1^2)$ component (component 1) and $L_i = 2$ if y_i is drawn from the $\mathrm{N}(\mu_2, \sigma_2^2)$ component (component 2). Then, the model can be written as

$$y_i \mid L_i, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2 \overset{ind.}{\sim} \mathrm{N}(y_i \mid \mu_{L_i}, \sigma_{L_i}^2)$$
$$\mathrm{P}(L_i = 1|w) = w = 1 - \mathrm{P}(L_i = 2|w)$$
$$(w, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) \sim p(w, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$$

- If we marginalize over L_i, for $i = 1, ..., n$, we recover the original mixture formulation.
- The inclusion of indicator variables is very common in finite mixture models, and it is also used extensively for DP mixtures.

Finite mixture models

- We can also write

$$w\mathrm{N}(y_i \mid \mu_1, \sigma_1^2) + (1-w)\mathrm{N}(y_i \mid \mu_2, \sigma_2^2) = \int \mathrm{N}(y_i \mid \mu, \sigma^2)\mathrm{d}G(\mu, \sigma^2),$$

  where

$$G = w\,\delta_{(\mu_1,\sigma_1^2)} + (1-w)\,\delta_{(\mu_2,\sigma_2^2)}$$

- A similar expression can be used for a general K mixture model.
- Note that G is discrete (and random) — a natural alternative is to use a DP prior for G, resulting in a Dirichlet process mixture (DPM) model, or more general nonparametric priors for discrete distributions.
- Working with a countable mixture (rather than a finite one) provides theoretical advantages (full support) as well as practical benefits: the number of mixture components is estimated from the data based on a model that supports a countable number of components in the prior.

Definition of the Dirichlet process mixture model

- The **Dirichlet process mixture model**

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

  where $K(\cdot \mid \theta)$ is a parametric distribution function indexed by θ.
- The Dirichlet process has been the most widely used prior for the random mixing distribution G, following the early work by Antoniak (1974), Lo (1984) and Ferguson (1983).
- Corresponding mixture density (or probability mass) function,

$$f(\cdot \mid G) = \int k(\cdot \mid \theta)\, \mathrm{d}G(\theta),$$

  where $k(\cdot \mid \theta)$ is the density (or probability mass) function of $K(\cdot \mid \theta)$.
- Because G is random, the c.d.f. $F(\cdot \mid G)$ and the density function $f(\cdot \mid G)$ are random (Bayesian nonparametric mixture models).

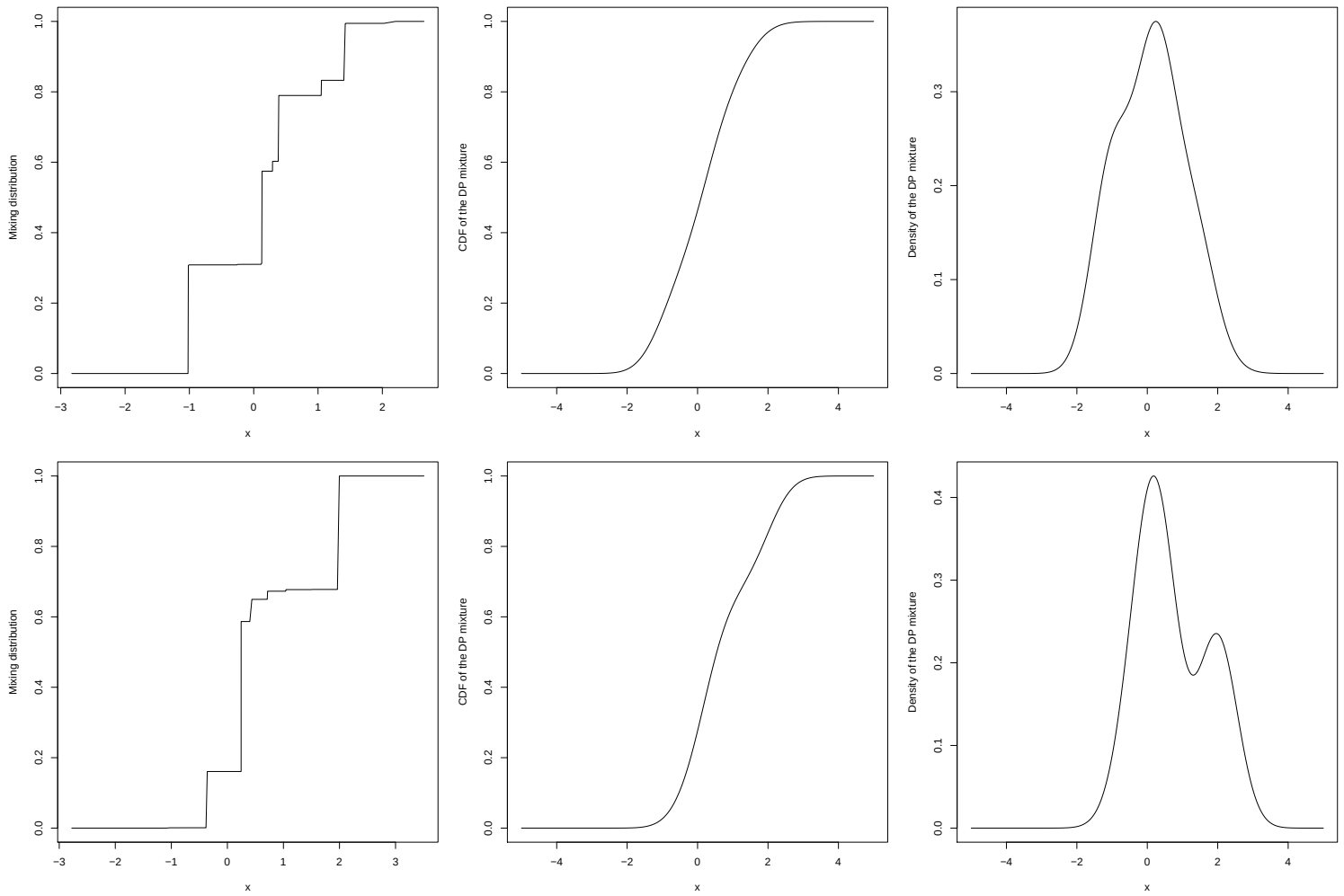


Figure 2.1: Two realizations from a DP($\alpha = 2$, $G_0 = \text{N}(0, 1)$) (left column) and the associated cumulative distribution function (center column) and density function (right column) for a location DP mixture of Gaussian kernels with standard deviation 0.6.

An equivalent formulation

- In the context of DP mixtures, the (almost sure) discreteness of realizations G from the DP(α, G_0) prior is an asset — it allows ties in the mixing parameters, and thus makes DP mixture models appealing for many applications, including density estimation and regression.

- Using the constructive definition of the DP, $G = \sum_{\ell=1}^{\infty} \omega_\ell \delta_{\vartheta_\ell}$, the prior probability model $f(\cdot \mid G)$ admits an (almost sure) representation as a countable mixture of parametric densities,

$$f(\cdot \mid G) = \sum_{\ell=1}^{\infty} \omega_\ell \, k(\cdot \mid \vartheta_\ell)$$

  - *Weights*: $\omega_1 = z_1$, $\omega_\ell = z_\ell \prod_{r=1}^{\ell-1}(1 - z_r)$, $\ell \geq 2$, with z_r i.i.d. Beta$(1, \alpha)$.
  - *Locations*: ϑ_ℓ i.i.d. G_0 (and the sequences $\{z_r : r = 1,2,\dots\}$ and $\{\vartheta_\ell : \ell = 1,2,\dots\}$ are independent).

Modeling options

- Contrary to DP prior models, DP mixtures can model
  - discrete distributions (e.g., $K(\cdot \mid \theta)$ might be Poisson or binomial)
  - and continuous distributions, either univariate ($K(\cdot \mid \theta)$ can be, e.g., normal, gamma, or uniform) or multivariate (with $K(\cdot \mid \theta)$, say, multivariate normal).

- Much more than just density estimation:
  - Non-Gaussian and non-linear regression through DP mixture modeling for the joint response-covariate distribution.
  - Flexible models for ordinal categorical responses.
  - Modeling of point process intensities through density estimation.
  - Time-series and/or spatial modeling, using dependent DP priors for temporally and/or spatially dependent mixing distributions.

Approximation or representation results for mixtures

- (Discrete) normal location-scale mixtures, $\sum_{j=1}^{M} w_j \, \mathrm{N}(\cdot \mid \mu_j, \sigma_j^2)$, can approximate arbitrarily well (as $M \to \infty$) any density on the real line (Ferguson, 1983; Lo, 1984).
- The c.d.f. of the Erlang mixture, $\sum_{j=1}^{J} w_j \, \mathrm{gamma}(t \mid j, \theta)$, converges pointwise to any continuous c.d.f. $H(t)$ on $\mathbb{R}^+$, as $J \to \infty$ and the common scale parameter $\theta \to 0$ (set $w_j = H(j\theta) - H((j-1)\theta)$).
- As $K \to \infty$, the Bernstein density, $\sum_{j=1}^{K} w_j \, \mathrm{Beta}(u \mid j, K - j + 1)$, converges uniformly to any continuous density $h(u)$ (with c.d.f. H) on $(0, 1)$ (set $w_j = H(j/K) - H((j-1)/K)$).
- For any non-increasing density $f(t)$ on the positive real line there exists a distribution function G such that f can be represented as a scale mixture of uniform densities: $f(t) = \int \theta^{-1} 1_{[0,\theta)}(t) \mathrm{d}G(\theta)$
  - The result yields flexible DP mixture models for symmetric unimodal densities (Brunner and Lo, 1989; Brunner, 1995) as well as general unimodal densities (Brunner, 1992; Lavine and Mockus, 1995; Kottas and Gelfand, 2001; Kottas and Krnjajić, 2009).

Support of Dirichlet process mixture models

- Results on Kullback-Leibler support for various types of DP mixture models (e.g., Wu and Ghosal, 2008).
- Consider the space of densities defined on sample space $\mathcal{X}$.
- For any density f_0 in that space, the Kullback-Leibler neighborhood of size $\varepsilon > 0$ is given by

$$K_\varepsilon(f_0) = \left\{ f : \int f_0(x) \log\left(\frac{f_0(x)}{f(x)}\right) \mathrm{d}x < \varepsilon \right\}$$

- A nonparametric prior model for densities satisfies the Kullback-Leibler property if it assignes positive probability to $K_\varepsilon(f_0)$ for any density f_0 in the space of interest, and for any $\varepsilon > 0$ (e.g., Walker, Damien and Lenk, 2004). Typically, some regularity conditions are needed for f_0.

Semiparametric Dirichlet process mixture models

- Typically, semiparametric DP mixtures are employed

$$y_i \mid G, \phi \overset{i.i.d.}{\sim} f(\cdot \mid G, \phi) = \int k(\cdot \mid \theta, \phi)\, \mathrm{d}G(\theta), \qquad i = 1, \ldots, n$$
$$G \sim \mathrm{DP}(\alpha, G_0)$$

  with a parametric prior $p(\phi)$ placed on ϕ (and, perhaps, hyperpriors for α and/or the parameters ψ of $G_0 \equiv G_0(\cdot \mid \psi)$).

- **Hierarchical formulation** for DP mixture models: introduce latent mixing parameter θ_i associated with y_i,

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

Parametric models in the two limits for α

- Two *limiting* special cases of the DP mixture model.
  - One distinct component, when $\alpha \to 0^+$

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

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

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

Connection with finite mixture models

- The countable sum formulation of the DP mixture model has motivated the study of several variants and extensions.
- It also provides a link between limits of finite mixtures, with prior for the weights given by a symmetric Dirichlet distribution, and DP mixture models (e.g., Ishwaran and Zarepour, 2000).
- Consider the finite mixture model with J components:

$$\sum_{j=1}^{J} q_j \, k(y \mid \vartheta_j),$$

with $(q_1, \ldots, q_J) \sim \text{Dir}(\alpha/J, \ldots, \alpha/J)$ and $\vartheta_j \overset{i.i.d.}{\sim} G_0$, $j = 1, \ldots, J$.

- When $J \to \infty$, this model corresponds to a DP mixture with kernel k and a DP(α, G_0) prior for the mixing distribution.
  - As $J \to \infty$, $\sum_{j=1}^{J} q_j \delta_{\vartheta_j}$ converges weakly to $\sum_{\ell=1}^{\infty} \omega_\ell \delta_{\vartheta_\ell} \sim \text{DP}(\alpha, G_0)$.

Prior specification

- Taking expectation over G with respect to its DP prior $\text{DP}(\alpha, G_0)$, we obtain:

$$\text{E}\{F(\cdot \mid G, \phi)\} = F(\cdot \mid G_0, \phi), \quad \text{E}\{f(\cdot \mid G, \phi)\} = f(\cdot \mid G_0, \phi).$$

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

- On the other hand, recall that for the $\text{DP}(\alpha, G_0)$, α controls how *close* a realization G is to G_0, but also the extent of discreteness of G.

- In the DP mixture model, α controls the prior distribution of the number of distinct elements n^* of vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, and hence the number of distinct mixture components associated with a sample of size n (Antoniak, 1974; Escobar and West, 1995; Liu, 1996).

Pólya urn revisited

- Consider the joint prior distribution for $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ that arises from the prior model for the mixing parameters, $\theta_i \mid G \overset{i.i.d.}{\sim} G$ with $G \mid \alpha, \psi \sim \text{DP}(\alpha, G_0(\cdot \mid \psi))$, after integrating G over its DP prior.

- Using the Pólya urn characterization of the DP,

$$p(\boldsymbol{\theta} \mid \alpha, \psi) = G_0(\theta_1 \mid \psi) \prod_{i=2}^{n} \left\{ \frac{\alpha}{\alpha + i - 1} G_0(\theta_i \mid \psi) + \frac{1}{\alpha + i - 1} \sum_{j=1}^{i-1} \delta_{\theta_j}(\theta_i) \right\}.$$

- As is essentially always the case for DP mixtures, assume that G_0 is a continuous distribution (i.e., it has no atoms) such that ties can only arise by setting θ_i equal to θ_j, for $j < i$.

Pólya urn revisited

- The prior distribution $p(\boldsymbol{\theta} \mid \alpha, \psi)$ can be written in an equivalent form which makes explicit the partitioning (clustering) induced by the discreteness of the DP prior (Antoniak, 1974; Lo, 1984).

- Denote by $\boldsymbol{\pi} = \{s_j : j = 1, ..., n^*\}$ a generic partition of $\{1, ..., n\}$, where: n^* is the number of cells of the partition; n_j is the number of elements in cell s_j; $e_{j,1} < ... < e_{j,n_j}$ are the elements of cell s_j.

- Letting $\mathcal{P}_n$ denote the set of all partitions of $\{1, ..., n\}$,

$$p(\boldsymbol{\theta} \mid \alpha, \psi) = \sum_{\boldsymbol{\pi} \in \mathcal{P}_n} p(\boldsymbol{\pi} \mid \alpha) \prod_{j=1}^{n^*} G_0(\theta_{e_{j,1}} \mid \psi) \prod_{i=2}^{n_j} \delta_{\theta_{e_{j,1}}}(\theta_{e_{j,i}})$$

where $p(\boldsymbol{\pi} \mid \alpha)$ is the DP induced prior probability for partition $\boldsymbol{\pi}$,

$$p(\boldsymbol{\pi} \mid \alpha) = \left(\prod_{m=1}^{n} (\alpha + m - 1) \right)^{-1} \alpha^{n^*} \prod_{j=1}^{n^*} (n_j - 1)!$$

Number of distinct components

- Prior expectation and variance for the number of distinct elements (partition cells), $n^* \equiv n^*(n)$, of vector $(\theta_1, \ldots, \theta_n)$.
- Let U_i, for $i = 1, ..., n$, be binary random variables with U_i indicating whether θ_i is a new value drawn from G_0 $(U_i = 1)$ or not $(U_i = 0)$.
- Conditional on α, the U_i are independent Bernoulli random variables with $\Pr(U_i = 1 \mid \alpha) = \alpha/(\alpha + i - 1)$, for $i = 1, ..., n$.
- Since $n^* = \sum_{i=1}^{n} U_i$, we obtain

$$\mathrm{E}(n^* \mid \alpha) = \sum_{i=1}^{n} \frac{\alpha}{\alpha + i - 1} \quad \text{and} \quad \mathrm{Var}(n^* \mid \alpha) = \sum_{i=1}^{n} \frac{\alpha(i-1)}{(\alpha + i - 1)^2}$$

- The prior moments for n^* can be used to guide the choice of the value for α, or the prior parameters for α. (The conditional expectation $\mathrm{E}(n^* \mid \alpha)$ can be averaged over the prior for α to obtain $\mathrm{E}(n^*)$.)

Number of distinct components

- A fairly accurate approximation (for practically all values of n and α):

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

  Hence, $\mathrm{E}(n^* \mid \alpha)$ increases at a logarithmic rate with n (for fixed α).
- Therefore, $\mathrm{E}(n^*(n) \mid \alpha) \to \infty$, as $n \to \infty$. In fact, $n^*(n)$ converges almost surely to ∞, as $n \to \infty$ (Korwar and Hollander, 1973).
  - Even though new distinct values are increasingly rare, the DP prior implies n^* which is steadily increasing with n.

- The full prior for the number of distinct elements can also be derived:

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

  where the factors $c_n(m) = \Pr(n^* = m \mid \alpha = 1)$ can be computed using certain recurrence formulas (Antoniak, 1974; Escobar and West, 1995; Ghosal and van der Vaart, 2017).
  - If α has prior $p(\alpha)$, $\Pr(n^* = m) = \int \Pr(n^* = m \mid \alpha) p(\alpha) \mathrm{d}\alpha$.

Methods for posterior inference

- Data $= \{y_i, i = 1, \ldots, n\}$ i.i.d., conditionally on G and ϕ, from $f(\cdot \mid G, \phi)$. (If the model includes a regression component, the data also include the covariate vectors $\mathbf{x}_i$, and, in such cases, ϕ, typically, includes the vector of regression coefficients).
- Interest in inference for the latent mixing parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, for ϕ (and the hyperparameters α, ψ), for $f(y_0 \mid G, \phi)$, and, in general, for functionals $H(F(\cdot \mid G, \phi))$ of the random mixture $F(\cdot \mid G, \phi)$ (e.g., c.d.f. function, hazard function, mean and variance functionals, percentile functionals).
- Full inference, given the data, for all these random quantities is based on the joint posterior distribution of the DP mixture model

$$p(G, \phi, \boldsymbol{\theta}, \alpha, \psi \mid \text{data})$$

Marginal posterior simulation methods

- The joint posterior distribution can be expressed as

$$p(G, \phi, \boldsymbol{\theta}, \alpha, \psi \mid \text{data}) = p(G \mid \boldsymbol{\theta}, \alpha, \psi) p(\boldsymbol{\theta}, \phi, \alpha, \psi \mid \text{data})$$

- $p(\boldsymbol{\theta}, \phi, \alpha, \psi \mid \text{data})$ is the marginal posterior for the finite-dimensional portion of the full *parameter vector* $(G, \phi, \boldsymbol{\theta}, \alpha, \psi)$.

- $G \mid \boldsymbol{\theta}, \alpha, \psi \sim \text{DP}(\tilde{\alpha}, \tilde{G}_0)$, where $\tilde{\alpha} = \alpha + n$, and

$$\tilde{G}_0(\cdot) = \frac{\alpha}{\alpha + n} G_0(\cdot \mid \psi) + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{\theta_i}(\cdot).$$

  (Hence, the c.d.f., $\tilde{G}_0(t) = \frac{\alpha}{\alpha+n} G_0(t \mid \psi) + \frac{1}{\alpha+n} \sum_{i=1}^{n} 1_{[\theta_i, \infty)}(t)$).

- Sampling from the $\text{DP}(\tilde{\alpha}, \tilde{G}_0)$ is possible using one of its definitions. We can thus obtain full posterior inference under DP mixture models if we sample from the marginal posterior $p(\boldsymbol{\theta}, \phi, \alpha, \psi \mid \text{data})$.

Marginal posterior simulation methods

- The marginal posterior $p(\boldsymbol{\theta}, \phi, \alpha, \psi \mid \text{data})$ corresponds to the marginalized version of the DP mixture model, obtained after integrating G over its DP prior (Blackwell and MacQueen, 1973),

$$\begin{aligned} y_i \mid \theta_i, \phi &\overset{ind.}{\sim} k(y_i \mid \theta_i, \phi), \qquad i = 1, \ldots, n \\ \boldsymbol{\theta} = (\theta_1, \ldots, \theta_n) \mid \alpha, \psi &\sim p(\boldsymbol{\theta} \mid \alpha, \psi), \\ \phi, \alpha, \psi &\sim p(\phi)p(\alpha)p(\psi). \end{aligned}$$

- The prior distribution $p(\boldsymbol{\theta} \mid \alpha, \psi)$ for the mixing parameters θ_i can be developed through the Pólya urn characterization of the DP,

$$p(\boldsymbol{\theta} \mid \alpha, \psi) = G_0(\theta_1 \mid \psi) \prod_{i=2}^{n} \left\{ \frac{\alpha}{\alpha + i - 1} G_0(\theta_i \mid \psi) + \frac{1}{\alpha + i - 1} \sum_{j=1}^{i-1} \delta_{\theta_j}(\theta_i) \right\}.$$

  Equivalently, the expression in terms of the DP induced partition structure can be used.

  - Either way, for increasing sample sizes, the joint prior $p(\boldsymbol{\theta} \mid \alpha, \psi)$ gets increasingly complex to work with.

Marginal posterior simulation methods

- Therefore, the marginal posterior

$$p(\boldsymbol{\theta}, \phi, \alpha, \psi \mid \text{data}) \propto p(\boldsymbol{\theta} \mid \alpha, \psi) p(\phi) p(\alpha) p(\psi) \prod_{i=1}^{n} k(y_i \mid \theta_i, \phi)$$

  is difficult to work with — even point estimates practically impossible to compute for moderate to large sample sizes.

- Early work for posterior inference:
  - Some results for certain problems in density estimation, i.e., expressions for Bayes point estimates of $f(y_0 \mid G)$ (e.g., Lo, 1984; Brunner and Lo, 1989).
  - Approximations for special cases, e.g., for binomial DP mixtures (Berry and Christensen, 1979).
  - Monte Carlo integration algorithms to obtain point estimates for the θ_i (Ferguson, 1983; Kuo, 1986a,b).

Simulation-based model fitting

- Note that, although the joint prior $p(\boldsymbol{\theta} \mid \alpha, \psi)$ has an awkward expression for samples of realistic size n, the prior full conditionals have convenient expressions:

$$p(\theta_i \mid \{\theta_j : j \neq i\}, \alpha, \psi) = \frac{\alpha}{\alpha + n - 1} G_0(\theta_i \mid \psi) + \frac{1}{\alpha + n - 1} \sum_{j \neq i} \delta_{\theta_j}(\theta_i)$$

- **Key idea** (Escobar, 1988; 1994): setup a Markov chain to explore the posterior $p(\boldsymbol{\theta}, \phi, \alpha, \psi \mid \text{data})$ by simulating only from posterior full conditional distributions, which arise by combining the likelihood terms with the corresponding prior full conditionals (in fact, Escobar's algorithm is essentially a Gibbs sampler developed for a specific class of models!).
- Several other Markov chain Monte Carlo (MCMC) methods that improve on the original algorithm (e.g., West et al., 1994; Escobar and West, 1995; Bush and MacEachern, 1996; Neal, 2000; Jain and Neal, 2004).

Simulation-based model fitting

- A key property for the implementation of the Gibbs sampler is the discreteness of G, which induces a partition (clustering) of the θ_i.
  - n^*: number of distinct elements (clusters) in the vector $(\theta_1, \ldots, \theta_n)$.
  - θ_j^*, $j = 1, \ldots, n^*$: the distinct θ_i.
  - $\boldsymbol{w} = (w_1, \ldots, w_n)$: vector of configuration indicators, defined by $w_i = j$ if and only if $\theta_i = \theta_j^*$, $i = 1, \ldots, n$.
  - n_j: size of j-th cluster, i.e., $n_j = |\{i : w_i = j\}|$, $j = 1, \ldots, n^*$.
- $(n^*, \boldsymbol{w}, (\theta_1^*, \ldots, \theta_{n^*}^*))$ is equivalent to $(\theta_1, \ldots, \theta_n)$.
- Standard Gibbs sampler to draw from $p(\boldsymbol{\theta}, \phi, \alpha, \psi \mid \text{data})$ (Escobar and West, 1995) is based on the following full conditionals:
  1. $p(\theta_i \mid \{\theta_{i'} : i' \neq i\}, \alpha, \psi, \phi, \text{data})$, for $i = 1, \ldots, n$.
  2. $p(\phi \mid \{\theta_i : i = 1, \ldots, n\}, \text{data})$.
  3. $p(\psi \mid \{\theta_j^* : i = 1, \ldots, n^*\}, n^*, \text{data})$.
  4. $p(\alpha \mid n^*, \text{data})$.

  (The expressions include conditioning only on the relevant variables, exploiting the conditional independence structure of the model and properties of the DP).

Simulation-based model fitting

1. For each $i = 1, \ldots, n$, $p(\theta_i \mid \{\theta_{i'} : i' \neq i\}, \alpha, \psi, \phi, \text{data})$ is simply a mixture of n^{*-} point masses and the posterior for θ_i based on y_i,

$$\frac{\alpha q_0}{\alpha q_0 + \sum_{j=1}^{n^{*-}} n_j^- q_j} h(\theta_i \mid \psi, \phi, y_i) + \sum_{j=1}^{n^{*-}} \frac{n_j^- q_j}{\alpha q_0 + \sum_{j=1}^{n^{*-}} n_j^- q_j} \delta_{\theta_j^{*-}}(\theta_i).$$

   - $q_j = k(y_i \mid \theta_j^{*-}, \phi)$
   - $q_0 = \int k(y_i \mid \theta, \phi) g_0(\theta \mid \psi) \mathrm{d}\theta$
   - $h(\theta_i \mid \psi, \phi, y_i) \propto k(y_i \mid \theta_i, \phi) g_0(\theta_i \mid \psi)$
   - g_0 is the density of G_0
   - The superscript "$-$" denotes all relevant quantities when θ_i is removed from the vector $(\theta_1, \ldots, \theta_n)$, e.g., n^{*-} is the number of clusters in $\{\theta_{i'} : i' \neq i\}$.
   - Updating θ_i implicitly updates w_i, $i = 1, \ldots, n$; before updating θ_{i+1}, we redefine n^*, θ_j^* for $j = 1, \ldots, n^*$, w_i for $i = 1, \ldots, n$, and n_j, for $j = 1, \ldots, n^*$.

Simulation-based model fitting

2. The posterior full conditional for ϕ does not involve the nonparametric part of the DP mixture model,

$$p(\phi \mid \{\theta_i : i = 1, \ldots, n\}, \text{data}) \propto p(\phi) \prod_{i=1}^{n} k(y_i \mid \theta_i, \phi).$$

3. Regarding the parameters ψ of G_0,

$$p(\psi \mid \{\theta_j^*, j = 1, \ldots, n^*\}, n^*, \text{data}) \propto p(\psi) \prod_{j=1}^{n^*} g_0(\theta_j^* \mid \psi),$$

leading to standard updates under a conditionally conjugate prior $p(\psi)$.

Simulation-based model fitting

4. Although the posterior full conditional for α is not of a standard form, an augmentation method facilitates sampling if α has a gamma prior (say, with mean a_α/b_α) (Escobar and West, 1995),

$$\begin{aligned} p(\alpha \mid n^*, \text{data}) &\propto p(\alpha)\alpha^{n^*}\frac{\Gamma(\alpha)}{\Gamma(\alpha+n)} \\ &\propto p(\alpha)\alpha^{n^*-1}(\alpha+n)\text{Beta}(\alpha+1,n) \\ &\propto p(\alpha)\alpha^{n^*-1}(\alpha+n)\int_0^1 \eta^\alpha(1-\eta)^{n-1}d\eta \end{aligned}$$

- Introduce an auxiliary variable η such that

$$p(\alpha,\eta \mid n^*, \text{data}) \propto p(\alpha)\,\alpha^{n^*-1}\,(\alpha+n)\,\eta^\alpha\,(1-\eta)^{n-1}$$

- Extend the Gibbs sampler to draw $\eta \mid \alpha, \text{data} \sim \text{Beta}(\alpha+1, n)$, and $\alpha \mid \eta, n^*, \text{data}$ from the two-component gamma mixture:

$$\epsilon\,\text{gamma}(a_\alpha+n^*, b_\alpha-\log(\eta))+(1-\epsilon)\,\text{gamma}(a_\alpha+n^*-1, b_\alpha-\log(\eta))$$

where $\epsilon = (a_\alpha+n^*-1)/\{n(b_\alpha-\log(\eta))+a_\alpha+n^*-1\}$.

Improved marginal Gibbs sampler

- (West et al., 1994; Bush and MacEachern, 1996): adds one more step where the cluster locations θ_j^* are resampled at each iteration to improve the mixing of the chain.
- At each iteration, once step (1) is completed, we obtain a specific number of clusters n^* and configuration $\boldsymbol{w} = (w_1, \ldots, w_n)$.
- After the marginalization over G, the prior for the θ_j^*, given the partition $(n^*, \boldsymbol{w})$, is given by $\prod_{j=1}^{n^*} g_0(\theta_j^* \mid \psi)$, i.e., given n^* and $\boldsymbol{w}$, the θ_j^* are i.i.d. from G_0.
- Hence, for each $j = 1, \ldots, n^*$, the posterior full conditional

$$p(\theta_j^* \mid \mathbf{w}, n^*, \psi, \phi, \text{data}) \propto g_0(\theta_j^* \mid \psi) \prod_{\{i: w_i = j\}} k(y_i \mid \theta_j^*, \phi).$$

More general marginal MCMC algorithms

- The Gibbs sampler can be difficult or inefficient to implement if:
  - The integral $\int k(y \mid \theta, \phi) g_0(\theta \mid \psi) \mathrm{d}\theta$ is not available in closed form (and numerical integration is not feasible or reliable).
  - Random generation from $h(\theta \mid \psi, \phi, y) \propto k(y \mid \theta, \phi) g_0(\theta \mid \psi)$ is not readily available.
- For such cases, alternative MCMC algorithms have been proposed in the literature (e.g., MacEachern and Müller, 1998; Neal, 2000; Dahl, 2005; Jain and Neal, 2007).
- Extensions for data structures that include missing or censored observations are also possible (Kuo and Smith, 1992; Kuo and Mallick, 1997; Kottas, 2006).

Posterior predictive distributions

- Implementing one of the available MCMC algorithms for DP mixture models, we obtain B posterior samples

$$\{\boldsymbol{\theta}_b = (\theta_{ib} : i = 1, \ldots, n), \alpha_b, \psi_b, \phi_b\}, \quad b = 1, \ldots, B,$$

from $p(\boldsymbol{\theta}, \phi, \alpha, \psi \mid \text{data})$.

- Or, equivalently, posterior samples

$$\left\{n_b^*, \mathbf{w}_b, \boldsymbol{\theta}_b^* = (\theta_{jb}^* : j = 1, \ldots, n_b^*), \alpha_b, \psi_b, \phi_b\right\}, \quad b = 1, \ldots, B,$$

from $p(n^*, \mathbf{w}, \boldsymbol{\theta}^* = (\theta_j^* : j = 1, \ldots, n^*), \phi, \alpha, \psi \mid \text{data})$.

- Bayesian *density estimate* is based on the posterior predictive density $p(y_0 \mid \text{data})$ corresponding to a *new* y_0 (with associated θ_0).

Posterior predictive distributions

- Using, again, the Pólya urn structure for the DP,

$$p(\theta_0 \mid n^*, \mathbf{w}, \boldsymbol{\theta}^*, \alpha, \psi) = \frac{\alpha}{\alpha + n} g_0(\theta_0 \mid \psi) + \frac{1}{\alpha + n} \sum_{j=1}^{n^*} n_j \delta_{\theta_j^*}(\theta_0).$$

- The posterior predictive density is given by

$$p(y_0 \mid \text{data}) = \int \int k(y_0 \mid \theta_0, \phi) p(\theta_0 \mid n^*, \mathbf{w}, \boldsymbol{\theta}^*, \alpha, \psi) \\ p(n^*, \mathbf{w}, \boldsymbol{\theta}^*, \alpha, \psi, \phi \mid \text{data}) \mathrm{d}\theta_0 \mathrm{d}\mathbf{w} \mathrm{d}\boldsymbol{\theta}^* \mathrm{d}\alpha \mathrm{d}\psi \mathrm{d}\phi$$

- Hence, a sample $\{y_{0,b} : b = 1, \ldots, B\}$ from the posterior predictive distribution can be obtained using the MCMC output, where, for each $b = 1, \ldots, B$:
  - we first draw $\theta_{0,b}$ from $p(\theta_0 \mid n_b^*, \mathbf{w}_b, \boldsymbol{\theta}_b^*, \alpha_b, \psi_b)$
  - and then, draw $y_{0,b}$ from $K(\cdot \mid \theta_{0,b}, \phi_b)$.

Posterior predictive distributions

- To further highlight the mixture structure, note that we can also write

$$p(y_0 \mid \text{data}) =$$

$$\int \left\{ \frac{\alpha}{\alpha + n} \int k(y_0 \mid \theta, \phi) g_0(\theta \mid \psi) \mathrm{d}\theta + \frac{n}{\alpha + n} \sum_{j=1}^{n^*} \frac{n_j}{n} k(y_0 \mid \theta_j^*, \phi) \right\}$$
$$p(n^*, \mathbf{w}, \boldsymbol{\theta}^*, \alpha, \psi, \phi \mid \text{data}) \mathrm{d}\mathbf{w} \mathrm{d}\boldsymbol{\theta}^* \mathrm{d}\alpha \mathrm{d}\psi \mathrm{d}\phi$$

- The integrand above is a mixture of:
  - the prior predictive density, $\mathrm{E}\{f(y_0 \mid G, \phi)\}$; and
  - a finite mixture with n^* components, with mixing parameters defined by the distinct θ_j^*, and weights given by n_j/n. This term dominates when α is small relative to n.
- The posterior predictive density for y_0 is obtained by averaging this mixture with respect to the posterior distribution of n^*, $\boldsymbol{w}$, $\boldsymbol{\theta}^*$ and all other parameters.

Inference for general functionals of the random mixture

- Note that $p(y_0 \mid \text{data})$ is the posterior point estimate for the density $f(y_0 \mid G, \phi)$ (at point y_0), i.e., $p(y_0 \mid \text{data}) = \mathrm{E}(f(y_0 \mid G, \phi) \mid \text{data})$.
  - The Bayesian density estimate under a DP mixture model can be obtained without sampling from the posterior distribution of G.
- Analogously, we can obtain posterior moments for $H(F(\cdot \mid G, \phi)) = \int H(K(\cdot \mid \theta, \phi))\mathrm{d}G(\theta)$, where H is a linear functional (Gelfand and Mukhopadhyay, 1995).
  - For linear functionals, the functional of the mixture is the mixture of the functionals applied to the parametric kernel (e.g., density and c.d.f. functionals, mean functional).
- How about more general types of inference?
  - Interval estimates for $F(y_0 \mid G, \phi)$ or $f(y_0 \mid G, \phi)$, for specified y_0?
  - Inference for non-linear functions of the c.d.f., e.g., cumulative hazard, $-\log(1 - F(y_0 \mid G, \phi))$, or hazard, $f(y_0 \mid G, \phi)/(1 - F(y_0 \mid G, \phi))$, functions?
  - Inference for other non-linear functionals, e.g., for percentiles?

Inference for general functionals of the random mixture

- Such inferences require the posterior distribution of G. Recall

$$p(G, \phi, \boldsymbol{\theta}, \alpha, \psi \mid \text{data}) = p(G \mid \boldsymbol{\theta}, \alpha, \psi) p(\boldsymbol{\theta}, \phi, \alpha, \psi \mid \text{data})$$

and

$$G \mid \boldsymbol{\theta}, \alpha, \psi \sim \text{DP}\left(\alpha + n, \tilde{G}_0(\cdot) = \frac{\alpha}{\alpha + n} G_0(\cdot \mid \psi) + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{\theta_i}(\cdot)\right)$$

- Hence, given posterior samples $(\boldsymbol{\theta}_b, \alpha_b, \psi_b, \phi_b)$, for $b = 1, \ldots, B$, from the marginalized version of the DP mixture, we can draw G_b from $p(G \mid \boldsymbol{\theta}_b, \alpha_b, \psi_b)$ using:
  - The original DP definition if we only need sample paths for the c.d.f. of the mixture (and y is univariate) (e.g., Krnjajić et al., 2008).
  - More generally, the DP constructive definition with a truncation approximation (Gelfand and Kottas, 2002; Ishwaran and Zarepour, 2002).

Inference for general functionals of the random mixture

- Applying directly the DP constructive definition,

$$G_b = \zeta_1 \, \delta_{U_1} + \sum_{\ell=2}^{L-1} \left\{ \zeta_\ell \prod_{r=1}^{\ell-1} (1-\zeta_r) \right\} \delta_{U_\ell} + \left\{ \prod_{r=1}^{L-1} (1-\zeta_r) \right\} \delta_{U_L}$$

where the ζ_ℓ, $\ell = 1, ..., L-1$, are i.i.d. Beta(1,$\alpha + n$), and (independently) the U_ℓ, $\ell = 1, ..., L$, are i.i.d. $\tilde{G}_0$.

- A more efficient truncation approximation through an alternative representation for the conditional posterior of G (Pitman, 1996)

$$G \mid (n^*, \boldsymbol{w}, \boldsymbol{\theta}^*), \alpha, \psi \stackrel{\mathcal{D}}{=} q_{n^*+1} G^* + \sum_{j=1}^{n^*} q_j \, \delta_{\theta_j^*}$$

where $G^* \mid \alpha, \psi \sim \text{DP}(\alpha, G_0(\psi))$ and, independently of G^*, the vector of weights, $(q_1, ..., q_{n^*}, q_{n^*+1}) \mid \alpha, \boldsymbol{w} \sim \text{Dirichlet}(n_1, ..., n_{n^*}, \alpha)$.

- Finally, the posterior samples G_b yield posterior samples $\{H(F(\cdot \mid G_b, \phi_b)) : b = 1, \ldots, B\}$ for any functional $H(F(\cdot \mid G, \phi))$.

Density estimation data example

- As an example, we analyze the `galaxy` data set: velocities (km/second) for 82 galaxies, drawn from six well-separated conic sections of the Corona Borealis region.
- The model is a location-scale DP mixture of Gaussian distributions, with a conjugate normal-inverse gamma baseline distribution:

$$f(y \mid G) = \int \mathrm{N}(y \mid \mu, \sigma^2)\, \mathrm{d}G(\mu, \sigma^2), \quad G \sim \mathrm{DP}(\alpha, G_0),$$

  where $G_0(\mu, \sigma^2) = \mathrm{N}(\mu \mid \mu_0, \sigma^2/\kappa)\mathrm{IGamma}(\sigma^2 \mid \nu, s)$.
- We consider four different prior specifications to explore the effect of increasing flexibility in the DP prior hyperparameters.
- Figure 2.2 shows posterior predictive density estimates obtained using the function `DPdensity` in the `R` package `DPpackage` (the code was taken from one of the examples in the help file).

Density estimation data example: Code

```
# Data     data(galaxy)
  galaxy = data.frame(galaxy,speeds=galaxy$speed/1000)
  attach(galaxy)
# Initial state
  state = NULL
# MCMC parameters
  nburn = 1000
  nsave = 10000
  nskip = 10
  ndisplay = 100
  mcmc = list(nburn=nburn,nsave=nsave,nskip=nskip,ndisplay=ndisplay)
# Example of Prior information 1
# Fixing alpha, m1, and Psi1
  prior1 = list(alpha=1,m1=rep(0,1),psiinv1=diag(0.5,1),nu1=4,tau1=1,tau2=100)
# Example of Prior information 2
# Fixing alpha and m1
  prior2 = list(alpha=1,m1=rep(0,1),psiinv2=solve(diag(0.5,1)),nu1=4,nu2=4,tau1=1,tau2=100)
# Example of Prior information 3
# Fixing only alpha
  prior3 = list(alpha=1,m2=rep(0,1),s2=diag(100000,1),psiinv2=solve(diag(0.5,1)),nu1=4,nu2=4,tau1=1,tau2=100)
# Example of Prior information 4
# Everything is random
  prior4 = list(a0=2,b0=1,m2=rep(0,1),s2=diag(100000,1),psiinv2=solve(diag(0.5,1)),nu1=4,nu2=4,tau1=1,tau2=100)
# Fit the models
  fit1.1 = DPdensity(y=speeds,prior=prior1,mcmc=mcmc,state=state,status=TRUE)
  fit1.2 = DPdensity(y=speeds,prior=prior2,mcmc=mcmc,state=state,status=TRUE)
  fit1.3 = DPdensity(y=speeds,prior=prior3,mcmc=mcmc,state=state,status=TRUE)
  fit1.4 = DPdensity(y=speeds,prior=prior4,mcmc=mcmc,state=state,status=TRUE)
# Plot the estimated density
  plot(fit1.1,ask=FALSE)
  plot(fit1.2,ask=FALSE)
  plot(fit1.3,ask=FALSE)
  plot(fit1.4,ask=FALSE)
```

Density estimation data example

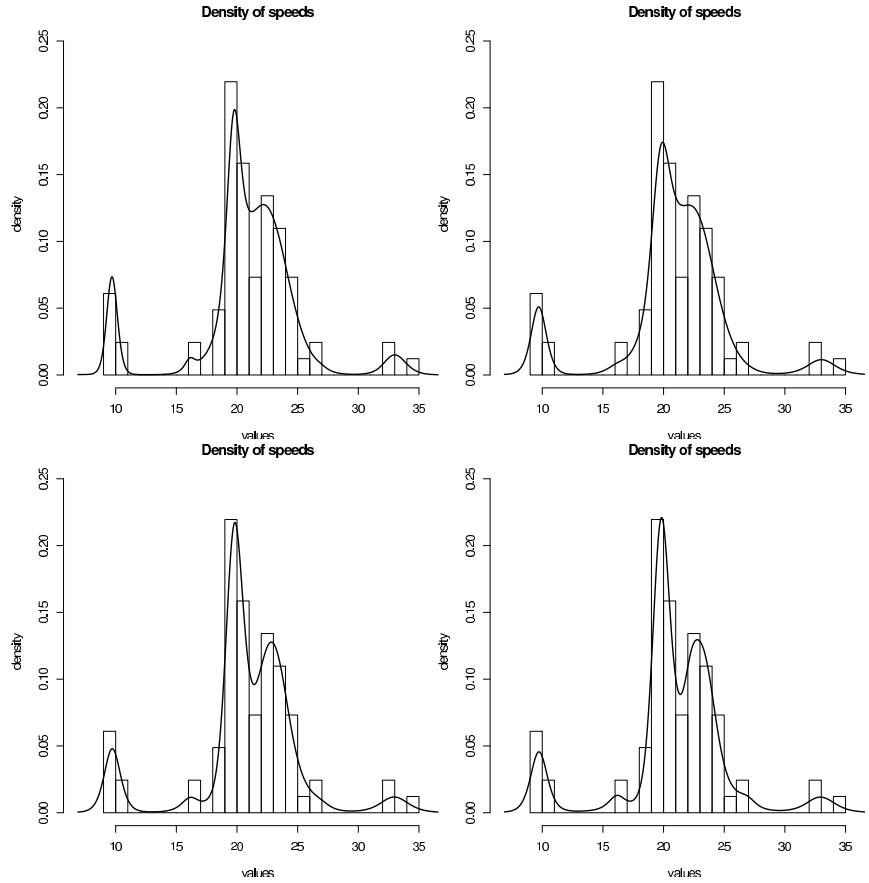


Figure 2.2: Histograms of the raw data and posterior predictive densities under four prior choices for the `galaxy` data. In the top left panel we set $\alpha = 1$, $\mu_0 = 0$, $s = 2$, $\nu = 4$, $\kappa \sim$ Gam(0.5, 50); the top right panel uses the same settings except $s \sim$ IGamma(4, 2); in the bottom left panel we add hyperprior $\mu_0 \sim$ N(0, 100000); and in the bottom right panel we further add hyperprior $\alpha \sim$ Gam(2, 2).

Conditional posterior simulation methods

- The main characteristic of the marginal MCMC methods is that they are based on the posterior distribution of the DP mixture model, $p(\boldsymbol{\theta}, \phi, \alpha, \psi \mid \text{data})$, resulting after marginalizing the random mixing distribution G (thus, referred to as *marginal* or *collapsed* methods).

- Although posterior inference for G is possible under the collapsed sampler, it is of interest to study alternative *conditional* posterior simulation approaches that impute G as part of the MCMC algorithm, and also improve on the mixing of marginal samplers.
  - Methods based on finite truncation approximation of G, using its stick-breaking representation – main example: Blocked Gibbs sampler (Ishwaran and Zarepour, 2000; Ishwaran and James, 2001).
  - Other approaches based on retrospective sampling techniques (Papaspiliopoulos and Roberts, 2008), slice sampling methods (Walker, 2007; Kalli et al., 2011), as well as combinations of retrospective and slice sampling (Yau et al., 2011).

Blocked Gibbs sampler

- Builds from truncation approximation to mixing distribution G given, for finite N, by

$$G_N = \sum_{\ell=1}^{N} p_\ell \, \delta_{Z_\ell}$$

  - The Z_ℓ, $\ell = 1, \ldots, N$, are i.i.d. G_0.
  - The weights arise through stick-breaking (with truncation)

$$p_1 = V_1, \quad p_\ell = V_\ell \prod_{r=1}^{\ell-1}(1 - V_r), \quad \ell = 2, \ldots, N-1, \quad p_N = \prod_{r=1}^{N-1}(1 - V_r),$$

    where the V_ℓ, $\ell = 1, \ldots, N-1$, are i.i.d. Beta$(1, \alpha)$.

- The joint prior for $\boldsymbol{p} = (p_1, \ldots, p_N)$, given α, corresponds to a special case of the generalized Dirichlet distribution (Connor and Mosimann, 1969),

$$f(\boldsymbol{p} \mid \alpha) = \alpha^{N-1} p_N^{\alpha-1}(1-p_1)^{-1}(1-(p_1+p_2))^{-1} \times \ldots \times (1 - \sum_{\ell=1}^{N-2} p_\ell)^{-1}.$$

The generalized Dirichlet distribution

- Assume that $V_\ell \overset{ind.}{\sim} \text{Beta}(a_\ell, b_\ell)$, for $\ell = 1, ..., N-1$, and define a probability vector, $\boldsymbol{p} = (p_1, ..., p_N)$, through

$$p_1 = V_1, \quad p_\ell = V_\ell \prod_{r=1}^{\ell-1}(1 - V_r), \quad \ell = 2, \ldots, N-1, \quad p_N = \prod_{r=1}^{N-1}(1 - V_r).$$

- Then, $\boldsymbol{p}$ follows a generalized Dirichlet distribution, with parameters $\boldsymbol{a} = (a_1, ..., a_{N-1})$ and $\boldsymbol{b} = (b_1, ..., b_{N-1})$, and with density given by

$$f(\boldsymbol{p} \mid \boldsymbol{a}, \boldsymbol{b}) = \left\{ \prod_{\ell=1}^{N-1} \frac{\Gamma(a_\ell + b_\ell)}{\Gamma(a_\ell)\Gamma(b_\ell)} \right\} p_1^{a_1 - 1} \times \ldots \times p_{N-1}^{a_{N-1}-1} p_N^{b_{N-1}-1} (1 - p_1)^{b_1 - (a_2 + b_2)}$$

$$(1 - (p_1 + p_2))^{b_2 - (a_3 + b_3)} \times \ldots \times \left(1 - \sum_{\ell=1}^{N-2} p_\ell\right)^{b_{N-2} - (a_{N-1} + b_{N-1})}$$

- If $b_{\ell-1} = a_\ell + b_\ell$, for $\ell = 2, ..., N-1$, the distribution reduces to a Dirichlet$(c_1, ..., c_N)$ with $c_\ell = a_\ell$, for $\ell = 1, ..., N-1$, and $c_N = b_{N-1}$.

Truncation level specification

- The DP truncation level N can be chosen to any desired level of accuracy.
- A simple approach based on the prior expectation for the partial sum of DP stick-breaking weights, $\mathrm{E}(\sum_{\ell=1}^{N} \omega_\ell \mid \alpha) = 1 - \{\alpha/(\alpha+1)\}^N$ (can be averaged over the prior for α to estimate $\mathrm{E}(\sum_{\ell=1}^{N} \omega_\ell)$).
  - For example, $\mathrm{E}(\sum_{\ell=1}^{25} \omega_\ell \mid \alpha = 2) = 0.99996$, and $\mathrm{E}(\sum_{\ell=1}^{75} \omega_\ell) = 0.99997$ under an exponential prior for α with mean 2.
- A more general approach, which involves also the sample size n, is available through Th. 2 in Ishwaran and James (2001): approximate upper bound of $4n\exp\{-(N-1)/\alpha\}$ on the L_1 distance between the prior predictive probability of the sample under the countable representation for G and its truncated version G_N.
  - For example, with $\alpha = 2$, the bound is 0.00001656 for $n = 10^2$ and $N = 35$, and it is 0.00001678 for $n = 10^7$ and $N = 58$.

Blocked Gibbs sampler

- Replacing G with $G_N \equiv (\boldsymbol{p}, \mathbf{Z})$, where $\mathbf{Z} = (Z_1, \ldots, Z_N)$, in the generic DP mixture model hierarchical formulation, we have:

$$y_i \mid \theta_i, \phi \overset{ind.}{\sim} k(y_i \mid \theta_i, \phi), \qquad i = 1, \ldots, n,$$

$$\theta_i \mid \mathbf{p}, \mathbf{Z} \overset{i.i.d.}{\sim} G_N, \qquad i = 1, \ldots, n,$$

$$\mathbf{p}, \mathbf{Z} \mid \alpha, \psi \sim f(\mathbf{p} \mid \alpha) \prod_{\ell=1}^{N} g_0(Z_\ell \mid \psi),$$

$$\phi, \alpha, \psi \sim p(\phi)p(\alpha)p(\psi).$$

- If we marginalize over the θ_i in the first two stages of the hierarchical model, we obtain a finite mixture model for the y_i,

$$f(y \mid \boldsymbol{p}, \mathbf{Z}, \phi) = \sum_{\ell=1}^{N} p_\ell \, k(y \mid Z_\ell, \phi)$$

(conditionally on $(\boldsymbol{p}, \mathbf{Z})$ and ϕ), which replaces the countable DP mixture, $f(y \mid G, \phi) = \int k(y \mid \theta, \phi) \, \mathrm{d}G(\theta) = \sum_{\ell=1}^{\infty} \omega_\ell \, k(y \mid \vartheta_\ell, \phi)$.

Blocked Gibbs sampler

- Now, having approximated the countable DP mixture with a finite mixture, the mixing parameters θ_i can be replaced with configuration variables $\mathbf{L} = (L_1, \ldots, L_n)$. Each L_i takes values in $\{1, \ldots, N\}$ such that $L_i = \ell$ if only if $\theta_i = Z_\ell$, for $i = 1, \ldots, n$ and $\ell = 1, \ldots, N$.
- Final version of the hierarchical model:

$$
\begin{aligned}
y_i \mid \mathbf{Z}, L_i, \phi &\overset{ind.}{\sim} k(y_i \mid Z_{L_i}, \phi), && i = 1, \ldots, n, \\
L_i \mid \mathbf{p} &\overset{i.i.d.}{\sim} \sum_{\ell=1}^{N} p_\ell \delta_\ell(L_i), && i = 1, \ldots, n, \\
Z_\ell \mid \psi &\overset{i.i.d.}{\sim} G_0(\cdot \mid \psi), && \ell = 1, \ldots, N, \\
\boldsymbol{p} \mid \alpha &\sim f(\boldsymbol{p} \mid \alpha), \\
\phi, \alpha, \psi &\sim p(\phi)p(\alpha)p(\psi).
\end{aligned}
$$

- Marginalizing over the L_i, we obtain the same finite mixture model for the y_i: $f(y \mid \boldsymbol{p}, \mathbf{Z}, \phi) = \sum_{\ell=1}^{N} p_\ell \, k(y \mid Z_\ell, \phi)$.

Posterior full conditional distributions

1. To update Z_ℓ for $\ell = 1, \ldots, N$:
   - Let n^* be the number of distinct values $\{L_j^* : j = 1, \ldots, n^*\}$ of vector $\mathbf{L}$.
   - Then, the posterior full conditional for Z_ℓ, $\ell = 1, \ldots, N$, can be expressed in general as:

$$p(Z_\ell \mid \ldots, \text{data}) \propto g_0(Z_\ell \mid \psi) \prod_{j=1}^{n^*} \prod_{\{i:L_i=L_j^*\}} k(y_i \mid Z_{L_j^*}, \phi)$$

   - If $\ell \notin \{L_j^* : j = 1, \ldots, n^*\}$, Z_ℓ is drawn from $G_0(\cdot \mid \psi)$
   - For $\ell = L_j^*$, $j = 1, \ldots, n^*$,

$$p(Z_{L_j^*} \mid \ldots, \text{data}) \propto g_0(Z_{L_j^*} \mid \psi) \prod_{\{i:L_i=L_j^*\}} k(y_i \mid Z_{L_j^*}, \phi)$$

2. The posterior full conditional for $\boldsymbol{p}$ is

$$p(\boldsymbol{p} \mid \ldots, \text{data}) \propto f(\boldsymbol{p} \mid \alpha) \prod_{\ell=1}^{N} p_\ell^{M_\ell},$$

where $M_\ell = |\{i : L_i = \ell\}|$, $\ell = 1, \ldots, N$.

- Results in a generalized Dirichlet distribution, which can be sampled through independent latent Beta variables.
- $V_\ell^* \overset{ind.}{\sim} \text{Beta}(1 + M_\ell, \alpha + \sum_{r=\ell+1}^{N} M_r)$, for $\ell = 1, \ldots, N-1$.
- $p_1 = V_1^*$; $p_\ell = V_\ell^* \prod_{r=1}^{\ell-1}(1 - V_r^*)$, for $\ell = 2, \ldots, N-1$; and $p_N = 1 - \sum_{\ell=1}^{N-1} p_\ell$.

3. Updating the L_i, $i = 1, \ldots, n$:

- Each L_i is drawn from the discrete distribution on $\{1, \ldots, N\}$ with probabilities $\tilde{p}_{\ell i} \propto p_\ell k(y_i \mid Z_\ell, \phi)$, for $\ell = 1, \ldots, N$.
- Note that the update for each L_i does not depend on the other $L_{i'}$, $i' \neq i$. This aspect of this Gibbs sampler, along with the *block updates* for the Z_ℓ, are key advantages over Pólya urn based marginal MCMC methods.

4. The posterior full conditional for ϕ is

$$p(\phi \mid \ldots, \text{data}) \propto p(\phi) \prod_{i=1}^{n} k(y_i \mid \theta_i, \phi).$$

5. The posterior full conditional for ψ is

$$p(\psi \mid \ldots, \text{data}) \propto p(\psi) \prod_{j=1}^{n^*} g_0(Z_{L_j^*} \mid \psi).$$

6. The posterior full conditional for α is proportional to $p(\alpha)\alpha^{N-1}p_N^{\alpha}$, which with a gamma$(a_\alpha, b_\alpha)$ prior for α, results in a gamma$(N+a_\alpha-1, b_\alpha-\log(p_N))$ distribution. (For numerical stability, compute $\log(p_N) = \log\prod_{r=1}^{N-1}(1-V_r^*) = \sum_{r=1}^{N-1}\log(1-V_r^*)$.)

Note that the posterior samples from $p(\mathbf{Z}, \boldsymbol{p}, \mathbf{L}, \phi, \alpha, \psi \mid \text{data})$ yield directly the posterior for G_N, and thus, full posterior inference for any functional of the (approximate) DP mixture $f(\cdot \mid G_N, \phi) \equiv f(\cdot \mid \mathbf{p}, \mathbf{Z}, \phi)$.

Posterior predictive inference

- Posterior predictive density for *new* y_0, with corresponding configuration variable L_0,

$$\begin{aligned} p(y_0 \mid \text{data}) &= \int k(y_0 \mid Z_{L_0}, \phi) \left(\sum_{\ell=1}^{N} p_\ell \delta_\ell(L_0) \right) \\ &\qquad p(\mathbf{Z}, \mathbf{p}, \mathbf{L}, \phi, \alpha, \psi \mid \text{data}) \mathrm{d}L_0 \, \mathrm{d}\mathbf{Z} \, \mathrm{d}\mathbf{L} \, \mathrm{d}\boldsymbol{p} \, \mathrm{d}\phi \, \mathrm{d}\alpha \, \mathrm{d}\psi \\ &= \int \left(\sum_{\ell=1}^{N} p_\ell k(y_0 \mid Z_\ell, \phi) \right) \\ &\qquad p(\mathbf{Z}, \boldsymbol{p}, \mathbf{L}, \phi, \alpha, \psi \mid \text{data}) \mathrm{d}\mathbf{Z} \, \mathrm{d}\mathbf{L} \, \mathrm{d}\boldsymbol{p} \, \mathrm{d}\phi \, \mathrm{d}\alpha \, \mathrm{d}\psi \\ &= \mathrm{E}(f(y_0 \mid \boldsymbol{p}, \mathbf{Z}, \phi) \mid \text{data}). \end{aligned}$$

- Hence, $p(y_0 \mid \text{data})$ can be estimated over a grid in y_0 by drawing samples $\{L_{0b} : b = 1, \ldots, B\}$ for L_0, based on the posterior samples for $\mathbf{p}$, and computing the Monte Carlo estimate

$$B^{-1} \sum_{b=1}^{B} k(y_0 \mid Z_{L_{0b}}, \phi_b),$$

where B is the posterior sample size.

Model checking/comparison for DP mixtures

- Posterior predictive estimation/sampling is straightforward for DP mixtures, and this allows using standard model checking/comparison techniques for (hierarchical) Bayesian models. Two examples are discussed next.

- Posterior predictive loss criterion (Gelfand and Ghosh, 1998): choose model that minimizes $D_k(M) = P(M) + \{k/(k+1)\}G(M)$, where:
  - $P(M) = \sum_{i=1}^{n} \text{Var}^{(M)}(y_{new,i} \mid \text{data})$ is a penalty term, and
  - $G(M) = \sum_{i=1}^{n} \{y_i - \text{E}^{(m)}(y_{new,i} \mid \text{data})\}^2$ is a goodness of fit term.
  - $\text{E}^{(M)}(y_{new,i} \mid \text{data})$ and $\text{Var}^{(M)}(y_{new,i} \mid \text{data})$ is the posterior predictive mean and posterior predictive variance under model M for replicated response $y_{new,i}$; in regression problems, the posterior predictive distribution for $y_{new,i}$ is evaluated for the observed vector of covariates $\boldsymbol{x}_i$.

  - $k \geq 0$ controls the weight assigned to the goodness of fit term.

Model checking/comparison for DP mixtures

- Conditional predictive ordinate (CPO) for observation y_i under model M: $\text{CPO}_i^{(M)} = p^{(M)}(y_i \mid \{y_j : j \neq i\})$, that is, the value of the posterior predictive density at y_i, given the data set excluding y_i.
  - Ratio $\text{CPO}_i^{(M_1)}/\text{CPO}_i^{(M_2)}$ describes how well model M_1 supports observation y_i relative to model M_2.
  - "Pseudo Bayes factor", $B_{12} = \prod_{i=1}^n (\text{CPO}_i^{(M_1)}/\text{CPO}_i^{(M_2)})$, is an aggregate summary of how well supported the data are by model M_1 relative to model M_2 (Geisser and Eddy, 1979).
  - "Log pseudo marginal likelihood" (LPML) for model M: $\text{LPML}_M = \log \prod_{i=1}^n \text{CPO}_i^{(M)}$, such that $B_{12} = \exp(\text{LPML}_{M_1} - \text{LPML}_{M_2})$.

- The Bayes factor requires the non-trivial computation of the DP mixture model marginal likelihood, $m(\boldsymbol{y})$, where $\boldsymbol{y} = (y_1, ..., y_n)$.
  - $m(\boldsymbol{y}) = \int L(\boldsymbol{y}; \phi, \alpha, \psi) p(\phi) p(\alpha) p(\psi) \, \mathrm{d}\phi \mathrm{d}\alpha \mathrm{d}\psi$
  - $L(\boldsymbol{y}; \phi, \alpha, \psi) = \int \{\prod_{i=1}^n k(y_i \mid \theta_i, \phi)\} p(\boldsymbol{\theta} \mid \alpha, \psi) \, \mathrm{d}\boldsymbol{\theta}$
  - One approach is given in Basu and Chib (2003), using sequential importance sampling to estimate the likelihood ordinate $L(\boldsymbol{y}; \phi, \alpha, \psi)$.

Alternative computational inference schemes

- Alternative (to MCMC) fitting techniques have been studied.
  - Sequential importance sampling (Liu, 1996; Quintana, 1998; MacEachern et al., 1999; Quintana and Newton, 2000; Carvalho et al., 2010).
  - Weighted Chinese restaurant algorithms (Ishwaran and Takahara, 2002; Ishwaran and James 2003).
  - Monte Carlo EM (Naskar and Das, 2004).
  - Predictive recursion (Newton and Zhang, 1999; Tokdar et al., 2009).
  - Variational algorithms (e.g., Blei and Jordan, 2006; Zobay, 2009).

- Posterior simulation for DP mixture models (and, more generally, Bayesian nonparametric models) for *large* datasets is an active area of research – some of the earlier contributions to scalable NPB methods include Guha (2010) and Wang and Dunson (2011).

Variational algorithms

- An alternative to MCMC methods which is very popular in the machine learning literature, and is gaining some traction within the statistics community; see Blei et al. (2017) for a recent review.

- Consider a generic model where $\boldsymbol{y}$ denotes the data and $\boldsymbol{\theta}$ collects all parameters. Variational algorithms aim at replacing the intractable posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ with a more tractable approximation $q_{\boldsymbol{\eta}}(\boldsymbol{\theta})$ whose parameters $\boldsymbol{\eta}$ are chosen to minimize

$$K(p||q) = \int \log\left(\frac{q_{\boldsymbol{\eta}}(\boldsymbol{\theta})}{p(\boldsymbol{\theta} \mid \boldsymbol{y})}\right) q_{\boldsymbol{\eta}}(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}$$

  the Kullback-Leibler divergence between $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ and $q_{\boldsymbol{\eta}}(\boldsymbol{\theta})$.

- Variational inference methods reformulate the problem of computing the posterior distribution as an easier (and faster!) to handle optimization problem. The main drawback is that, in contrast to MCMC methods, there is no general theory that ensures convergence to the posterior distribution.

Variational algorithms

- The minimization of $K(p||q)$ can be alternatively approached as maximization of a lower bound on the log marginal likelihood:

$$\log p(\boldsymbol{y}) \geq \mathrm{E}_q \left\{\log p(\boldsymbol{\theta}, \boldsymbol{y})\right\} - \mathrm{E}_q \left\{\log q_{\boldsymbol{\eta}}(\boldsymbol{\theta})\right\}.$$

  The gap in the bound is the K-L divergence between $q_{\boldsymbol{\eta}}$ and the true posterior, and can be used to select among multiple solutions.

- Two key ingredients for an efficient algorithm: the particular form of the variational distribution $q_{\boldsymbol{\eta}}$, and the optimization procedure.
- In principle, there is a lot of freedom in choosing $q_{\boldsymbol{\eta}}$. Practical limitations arise from the need to have a tractable approximation and to compute the expectations $\mathrm{E}_q \left\{\log q_{\boldsymbol{\eta}}(\boldsymbol{\theta})\right\}$ and $\mathrm{E}_q \left\{\log p(\boldsymbol{\theta} \mid \boldsymbol{y})\right\}$.

- Variational approximations are relatively straightforward to work with in conditionally conjugate models (the same type of models for which Gibbs sampling is well suited). Extensions for non-conjugate models are more challenging.

Mean-field variational algorithms

- *Mean-field* methods use a factorized variational distribution $q_{\boldsymbol{\eta}}(\boldsymbol{\theta}) = \prod_{k=1}^{K} q_{k,\boldsymbol{\eta}_k}(\theta_k)$, where $\boldsymbol{\theta} = (\theta_1, ..., \theta_K)$, and are very popular due to their tractability.
- In particular, if the conditional posterior distribution for each θ_k belongs to the exponential family, it is natural to select $q_{k,\boldsymbol{\eta}_k}(\theta_k)$ as a member of the conditionally conjugate prior family.
  - Then, the optimization of the K-L divergence w.r.t. a single variational parameter $\boldsymbol{\eta}_k$ is achieved by computing the expectation (w.r.t. $q_{\boldsymbol{\eta}}$) of the exponential family natural parameter for θ_k.
  - Recursively updating each $\boldsymbol{\eta}_k$ by computing this expectation corresponds to performing coordinate ascent in the K-L divergence.
  - Hence, the algorithm has the flavor of a Gibbs sampler.
- In general, (mean-field) variational algorithms tend to underestimate the uncertainty in the posterior distribution and get the dependence among parameters wrong.

Mean-field variational methods for DP mixtures

- An early example of mean-field methods for DP mixture models with exponential family kernels and the corresponding conjugate prior as the DP centering distribution (Blei and Jordan, 2006).
- Consider the same model formulation with the blocked Gibbs sampler (although here the truncation is applied only to the variational distribution) and focus on parameters: $\{V_\ell : \ell = 1, ..., N-1\}$ that define the mixture weights through stick-breaking; the atoms $\{Z_\ell : \ell = 1, ..., N\}$; and the configuration variables $\{L_i : i = 1, ..., n\}$.
- Then, the variational distribution is given by

$$\prod_{\ell=1}^{N-1} q_{\gamma_\ell}(V_\ell) \prod_{\ell=1}^{N} q_{\psi_\ell}(Z_\ell) \prod_{i=1}^{n} q_{\xi_i}(L_i)$$

  where $q_{\gamma_\ell}(V_\ell) = \text{Beta}(V_\ell \mid \gamma_{\ell,1}, \gamma_{\ell,2})$, $q_{\psi_\ell}(Z_\ell) = G_0(Z_\ell \mid \psi_\ell)$, and $q_{\xi_i}(L_i) = \text{Mult}(\xi_i)$.

- Main inference focus on posterior predictive estimation – inference for more general functionals becomes more computationally intensive.

Applications of DP mixture models: some references

Dirichlet process mixture models, and their extensions, have largely dominated applied Bayesian nonparametric work, after the technology for their simulation-based model fitting was introduced. Included below is a sample of references categorized by methodological/application area.

- Density estimation, mixture deconvolution, and density regression: West et al. (1994); Escobar and West (1995); Cao and West (1996); Gasparini (1996); Müller et al. (1996); Ishwaran and James (2002); Do, Müller and Tang (2005); Leslie et al. (2007); Lijoi, Mena and Prünster (2007); Taddy and Kottas (2010).

- Generalized linear, and linear mixed, models; methods for longitudinal data analysis: Bush and MacEachern (1996); Kleinman and Ibrahim (1998a,b); Mukhopadhyay and Gelfand (1997); Müller and Rosner (1997); Quintana (1998); Kyung, Gill and Casella (2010); Hannah et al. (2011); Quintana et al. (2016).

Applications of DP mixture models: some references

- Methods for longitudinal cluster analysis and for functional clustering: Ray and Mallick (2006); Bigelow and Dunson (2009); Petrone, Guindani and Gelfand (2009).
- Regression modeling with structured error distributions and/or regression functions: Brunner (1995); Lavine and Mockus (1995); Kottas and Gelfand (2001); Dunson (2005); Kottas and Krnjajić (2009).
- Regression models for survival/reliability data: Kuo and Mallick (1997); Gelfand and Kottas (2003); Merrick et al. (2003); Hanson (2006); Argiento et al. (2009); De Iorio et al. (2009).
- Models for binary and ordinal data: Basu and Mukhopadhyay (2000); Hoff (2005); Das and Chattopadhyay (2004); Kottas, Müller and Quintana (2005); Shahbaba and Neal (2009); Bao and Hanson (2015); DeYoreo and Kottas (2015, 2018a,b).

Applications of DP mixture models: some references

- Errors-in-variables models; multiple comparisons problems; analysis of selection models: Müller and Roeder (1997); Gopalan and Berry (1998); Lee and Berger (1999).
- ROC data analysis: Erkanli et al. (2006); Hanson, Kottas and Branscum (2008).
- Meta-analysis and nonparametric ANOVA models: Mallick and Walker (1997); Tomlinson and Escobar (1999); Burr et al. (2003); De Iorio et al. (2004); Müller et al. (2004); Müller et al. (2005).
- Mixture models for Markov time series; time series modeling and econometrics applications: Müller, West and MacEachern (1997); Chib and Hamilton (2002); Hirano (2002); Hasegawa and Kozumi (2003); Griffin and Steel (2004); Tang and Ghosal (2007); Di Lucca et al. (2013); Antoniano-Villalobos and Walker (2016); DeYoreo and Kottas (2017); Kalli and Griffin (2018).

Semiparametric random effects models

- Linear random effects models (e.g., Laird and Ware, 1982) are a widely used class of models for repeated measurements,

$$\boldsymbol{y}_i = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i, \qquad i = 1, \ldots, n,$$

  where: $\boldsymbol{y}_i$ is the response vector for the i-th subject; $\boldsymbol{\beta}$ is the vector of fixed effects regression parameters; $\boldsymbol{b}_i$ is the vector of random effects; X_i and Z_i are covariate matrices associated with the fixed and random effects, respectively; and $\boldsymbol{\epsilon}_i$ is the vector of observational errors.

- It is common to assume that $\boldsymbol{b}_i$ is independent from $\boldsymbol{\epsilon}_i$, and that $\boldsymbol{\epsilon}_i \sim \mathrm{N}(0, \sigma^2 I)$.

- Furthermore, it is very common to assume that $\boldsymbol{b}_i \sim \mathrm{N}(0, D)$, mostly because of computational convenience.

Semiparametric random effects models

- Consider a special case, the random intercepts model:

$$y_{ij} = \mu + \theta_i + \epsilon_{ij}, \qquad \theta_i \sim \mathsf{N}(0, \tau^2), \qquad \epsilon_{ij} \sim \mathsf{N}(0, \sigma^2),$$

  for $j = 1, \ldots, m_i$ and $i = 1, \ldots, n$.

- A Bayesian formulation of this model also includes priors on μ, τ^2 and σ^2, e.g,

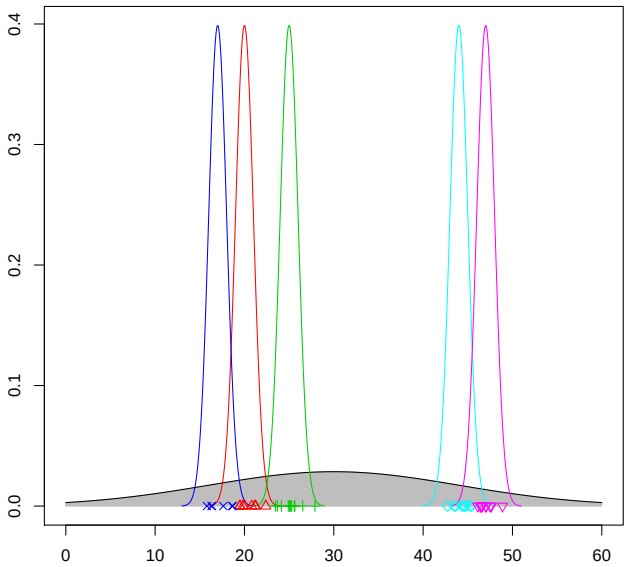
$$\mu \sim \mathsf{N}(\mu_0, \kappa^2) \qquad \sigma^2 \sim \mathsf{IG}(a, b) \qquad \tau^2 \sim \mathsf{IG}(c, d)$$

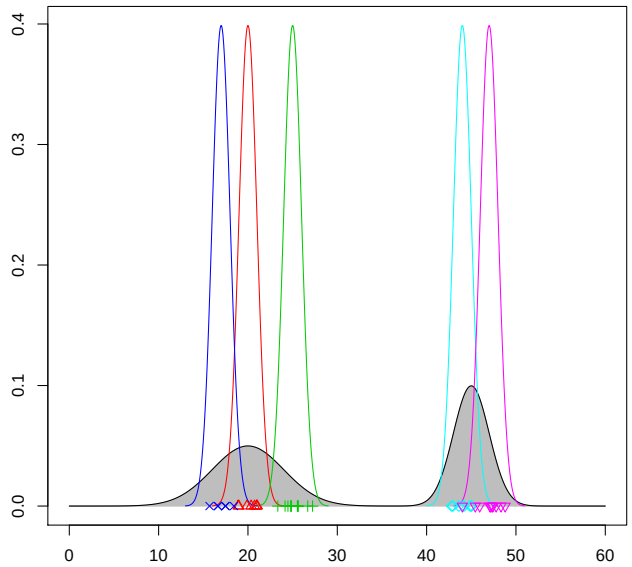
  (When selecting hyperparameters, recall that an improper prior for σ^2 would be OK, but improper priors for τ^2 are not.)

- When is the assumption of normality for the random effects distribution reasonable?

Random effects distributions

- Normality is, in general, an inappropriate assumption for the random effects distribution.





- Instead, we would often expect the random effects distribution to present multimodalities because of the effects of covariates that have not been included in the model.

Bayesian semiparametric random effects models

- Bayesian semiparametric random effects models have been discussed in Bush and MacEachern (1996), Kleinman and Ibrahim (1998a,b), Mukhopadhyay and Gelfand (1997), Burr and Doss (2005), and Kyung, Gill and Casella (2010), in addition to a number of applied papers.

- General formulation:

$$\begin{aligned} \boldsymbol{y}_i \mid \boldsymbol{\beta}, \boldsymbol{b}_i, \sigma^2 &\sim \mathrm{N}(X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i, \sigma^2 I), && i = 1, \ldots, n \\ \boldsymbol{b}_i \mid G &\sim G, && i = 1, \ldots, n \\ G \mid \alpha, D &\sim \mathrm{DP}(\alpha, \mathrm{N}(0, D)) \\ \boldsymbol{\beta}, \sigma^2, \alpha, D &\sim p(\beta, \sigma^2, \alpha, D) \end{aligned}$$

Bayesian semiparametric random intercepts model

- For the random intercepts model:

$$
\begin{aligned}
y_{ij} \mid \theta_i, \sigma^2 &\sim \mathrm{N}(\theta_i, \sigma^2), && j = 1, \ldots, m_i, \quad i = 1, \ldots, n \\
\theta_i \mid G &\sim G, && i = 1, \ldots, n \\
G \mid \alpha, \mu, \tau^2 &\sim \mathrm{DP}(\alpha, \mathrm{N}(\mu, \tau^2))
\end{aligned}
$$

  with hyperpriors for σ^2 and (some of) the DP parameters (α, μ, τ^2) (note that, without loss of generality, we absorbed the intercept μ).

- For $\alpha \to \infty$ we recover the traditional Gaussian random effects model, whereas for $\alpha \to 0$, the model reduces to a parametric model without random effects.
- For values of α in between, the model induces ties among the θ_i.

Fitting linear mixed models in R

- The DPpackage includes functions to fit (generalized) linear mixed models in which the random effects distribution is assigned a DP prior.
- We illustrate with a linear mixed model (function DPlmm).
- Data corresponds to growth information of 20 preadolescent schoolgirls reported by Goldstein (1979, Table 4.3, p. 101). Four variables are included:
  - height: a numeric vector giving the height in cm.
  - child: an ordered factor giving a unique identifier for the subject in the study.
  - age: a numeric vector giving the age of the child in years.
  - group: a factor with levels 1 (short), 2 (medium), and 3 (tall) giving the mother category.
- The height of girls was measured on a yearly basis from age 6 to 10. The measurements are given at exact years of age.

Fitting linear mixed models in R

- We fit the model

$$y_{ij} \mid (\theta_i, \beta_i), \sigma^2 \sim \mathrm{N}\left(\theta_i + x_{ij}\beta_i, \sigma^2\right)$$
$$(\theta_i, \beta_i) \mid G \sim \mathrm{DP}\left(\alpha, \mathrm{N}(\boldsymbol{\mu}, \Sigma)\right)$$

  where
  - y_{ij} is the j-th height observation for the i-th child.
  - x_{ij} is the age associated with the j-th height observation for the i-th child.
- Note that we do not use the variable `group` in the analysis (it may be used to add another layer to the hierarchy).
- Note the syntax of the function call (similar to other functions for random effects models in R):

```
DPlmm(fixed=height ~ 1, random= ~ age|child,
prior=prior, mcmc=mcmc, state=state, status=FALSE)
```

Fitting linear mixed models in R

- This R code is provided with the help file of DPlmm.

```
# School Girls Data Example
 data(schoolgirls)
 attach(schoolgirls)
# Prior information
 prior=list(alpha=1,nu0=4.01,tau1=0.01,tau2=0.01,tinv=diag(10,2),mub=rep(0,2),Sb=diag(1000,2))
# Initial state
 state = NULL
# MCMC parameters
 nburn=5000
 nsave=40000
 nskip=20
 ndisplay=1000
 mcmc = list(nburn=nburn,nsave=nsave,nskip=nskip,ndisplay=ndisplay)
# Fit the model:  First run
 fit1=DPlmm(fixed=height ~ 1,random= ~ age|child,prior=prior,mcmc=mcmc,state=state,status=TRUE)
 fit1
# Fit the model:  Continuation
 state=fit1$state
 fit2=DPlmm(fixed=height ~ 1,random= ~ age|child,prior=prior,mcmc=mcmc,state=state,status=FALSE)
 fit2
# Summary with HPD and Credibility intervals
 summary(fit2)
 summary(fit2,hpd=FALSE)
# Extract expected means of the random effect coefficients
 DPrandom(fit2)
# Plot an specific model parameter
 quartz()
 plot(fit2,ask=FALSE,nfigr=1,nfigc=2,param="sigma-(Intercept)")
 quartz()
 plot(fit2,ask=FALSE,nfigr=1,nfigc=2,param="ncluster")
```

Fitting linear mixed models in R

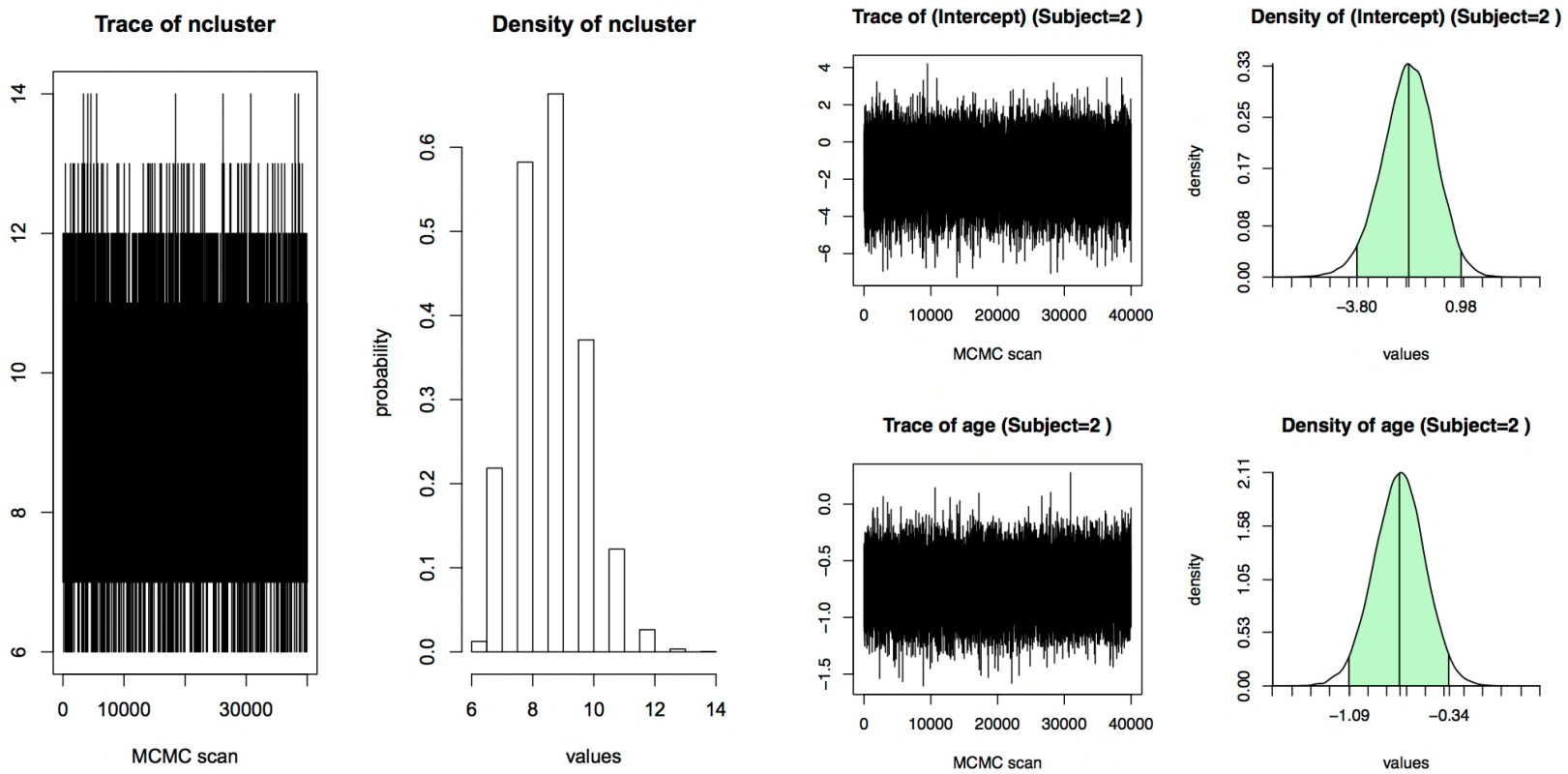


Figure 2.3. On the left, trace plot and histogram for the number of clusters generated by `DPlmm` in the school girls example. On the right, trace and posterior density plots for the parameters of one of the individuals.

Density regression using Dirichlet process mixtures

- Two dominant trends in the Bayesian regression literature: seek increasingly flexible regression function models, and accompany these models with general error distributions.
- Typically, Bayesian nonparametric modeling focuses on either the regression function or the error distribution.
- Bayesian nonparametric models for *density regression* (aka *conditional regression*) (West et al., 1994; Müller et al., 1996).
  - Flexible nonparametric mixture modeling for the joint distribution of response(s) and covariates.
  - Inference for the conditional response distribution given covariates.
- Both the response distribution and, implicitly, the regression relationship are modeled nonparametrically, thus providing a flexible framework for the general regression problem.

Density regression using Dirichlet process mixtures

- Focus on univariate continuous response y (though extensions for categorical and/or multivariate responses also possible).
- DP mixture model for the joint density $f(y, \mathbf{x})$ of the response y and the vector of covariates $\mathbf{x}$:

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

- For the mixture kernel $k(y, \mathbf{x} \mid \boldsymbol{\theta})$ use:
  - Multivariate normal for ($\mathbb{R}$-valued) continuous response and covariates.
  - Mixed continuous/discrete distribution to incorporate both categorical and continuous covariates.
  - Kernel component for y supported by $\mathbb{R}^+$ for problems in survival/reliability analysis.

Density regression using Dirichlet process mixtures

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

- Key features of the modeling approach:
  - Model for both non-linear regression curves **and** non-standard shapes for the conditional response density.
  - Model does not rely on additive regression formulations; it can uncover interactions between covariates that might influence the regression relationship.

Mean regression functional under normal DP mixtures

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

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

- Consider the decomposition of $\boldsymbol{\mu}_\ell = (\mu_\ell^y, \boldsymbol{\mu}_\ell^{\boldsymbol{x}})$ and $\Sigma_\ell = (\Sigma_\ell^y, \Sigma_\ell^{y\boldsymbol{x}}, \Sigma_\ell^{\boldsymbol{x}})$ into components that correspond to the response and covariates.
- Then, $f(y \mid \boldsymbol{x}, G) = \sum_{\ell=1}^{\infty} q_\ell(\boldsymbol{x}) \, \mathrm{N}(y \mid \lambda_\ell(\boldsymbol{x}), \tau_\ell^2)$, where
  - $q_\ell(\boldsymbol{x}) = \omega_\ell \mathrm{N}_p(\boldsymbol{x} \mid \boldsymbol{\mu}_\ell^{\boldsymbol{x}}, \Sigma_\ell^{\boldsymbol{x}}) / \{\sum_{s=1}^{\infty} \omega_s \mathrm{N}_p(\boldsymbol{x} \mid \boldsymbol{\mu}_s^{\boldsymbol{x}}, \Sigma_s^{\boldsymbol{x}})\}$
  - $\lambda_\ell(\boldsymbol{x}) = \mu_\ell^y + \Sigma_\ell^{y\boldsymbol{x}} (\Sigma_\ell^{\boldsymbol{x}})^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_\ell^{\boldsymbol{x}})$ and $\tau_\ell^2 = \Sigma_\ell^y - \Sigma_\ell^{y\boldsymbol{x}} (\Sigma_\ell^{\boldsymbol{x}})^{-1} (\Sigma_\ell^{y\boldsymbol{x}})^T$
- Mean regression function:

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

where $\beta_{0\ell} = \mu_\ell^y - \Sigma_\ell^{y\boldsymbol{x}} (\Sigma_\ell^{\boldsymbol{x}})^{-1} \boldsymbol{\mu}_\ell^{\boldsymbol{x}}$, and $\beta_{r\ell}$, for $r = 1, ..., p$, are the elements of vector $\Sigma_\ell^{y\boldsymbol{x}} (\Sigma_\ell^{\boldsymbol{x}})^{-1}$.

Synthetic data example

- Simulated data set with a continuous response y, one continuous covariate x_c, and one binary categorical covariate x_d.
  - x_{ci} independent $\mathrm{N}(0, 1)$.
  - $x_{di} \mid x_{ci}$ independent $\mathrm{Ber}(\mathrm{probit}(x_{ci}))$.
  - $y_i \mid x_{ci}, x_{di}$ ind. $\mathrm{N}(h(x_{ci}), \sigma_{x_{di}})$, with $\sigma_0 = 0.25$, $\sigma_1 = 0.5$, and

$$h(x_c) = 0.4x_c + 0.5\sin(2.7x_c) + 1.1(1 + x_c^2)^{-1}.$$

- Two sample sizes: $n = 200$ and $n = 2000$.

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

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

with

$$G \sim \mathrm{DP}(\alpha, G_0(\boldsymbol{\mu}, \Sigma, \pi) = \mathrm{N}_2(\boldsymbol{\mu}; \mathbf{m}, V)\mathrm{IW}(\Sigma; \nu, S)\mathrm{Beta}(\pi; a, b)).$$

Synthetic data example

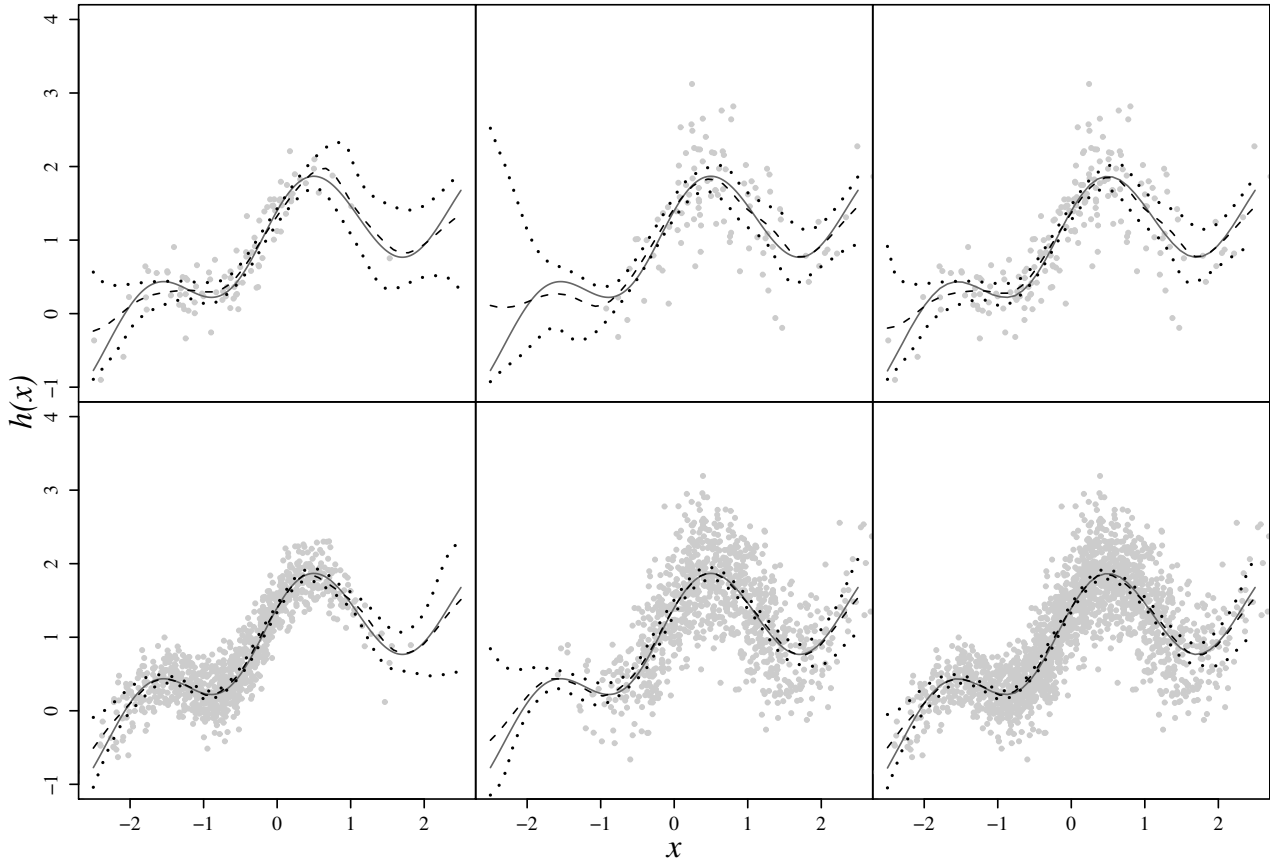


Figure 2.4. Posterior point and 90% interval estimates (dashed and dotted lines) for conditional response expectation $\mathrm{E}(y \mid x_c, x_d = 0; G)$ (left panels), $\mathrm{E}(y \mid x_c, x_d = 1; G)$ (middle panels), and $\mathrm{E}(y \mid x_c; G)$ (right panels). The corresponding data is plotted in grey for the sample of size $n = 200$ (top panels) and $n = 2000$ (bottom panels). The solid line denotes the true curve.

Quantile regression

- In regression settings, the covariates may have effect not only on the location of the response distribution but also on its shape.

- Model-based nonparametric approach to quantile regression.
  - Model joint density $f(y, \boldsymbol{x})$ of the response y and the M-variate vector of (continuous) covariates $\boldsymbol{x}$ with a DP mixture of normals:

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

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

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

- Key features of the DP mixture modeling framework:
  - Enables simultaneous inference for more than one quantile regression.
  - Allows flexible response distributions **and** non-linear quantile regression relationships.

Quantile regression: data example

- *Moral hazard* data on the relationship between shareholder concentration and several indices for managerial moral hazard in the form of expenditure with scope for private benefit (Yafeh & Yoshua, 2003).
  - Data set includes a variety of variables describing 185 Japanese industrial chemical firms listed on the Tokyo stock exchange.
  - Response y: index *MH5*, consisting of general sales and administrative expenses deflated by sales.
  - Four-dimensional covariate vector $\mathbf{x}$: *Leverage* (ratio of debt to total assets); log(*Assets*); *Age* of the firm; and *TOPTEN* (the percent of ownership held by the ten largest shareholders).

Quantile regression: data example

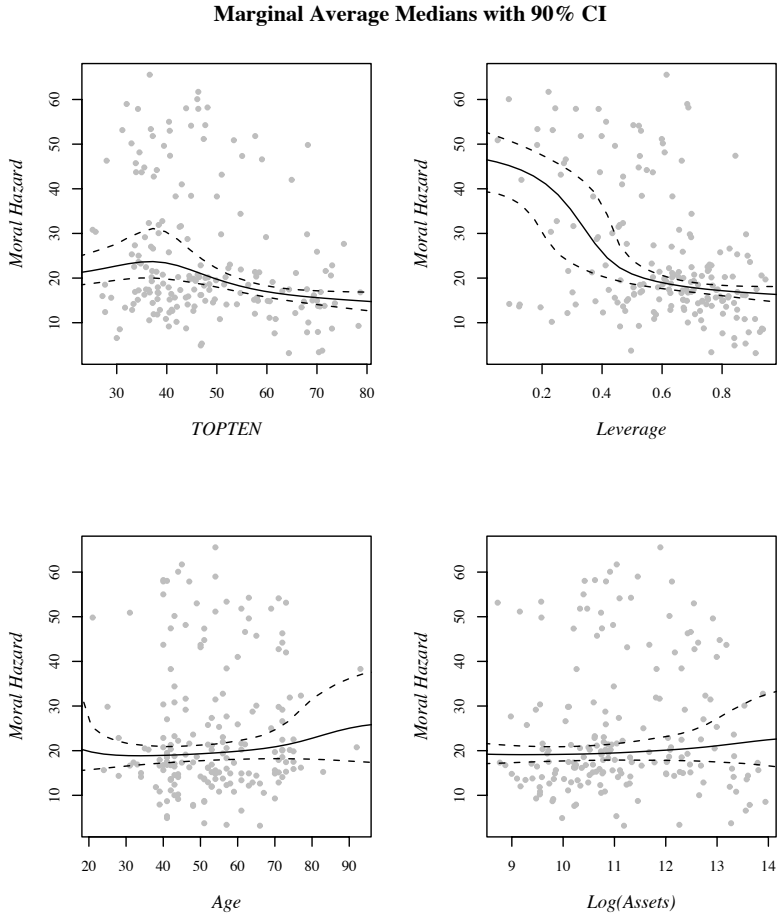


Figure 2.5. Posterior mean and 90% interval estimates for median regression for *MH*5 conditional on each individual covariate. Data scatterplots are shown in grey.

Quantile regression: data example

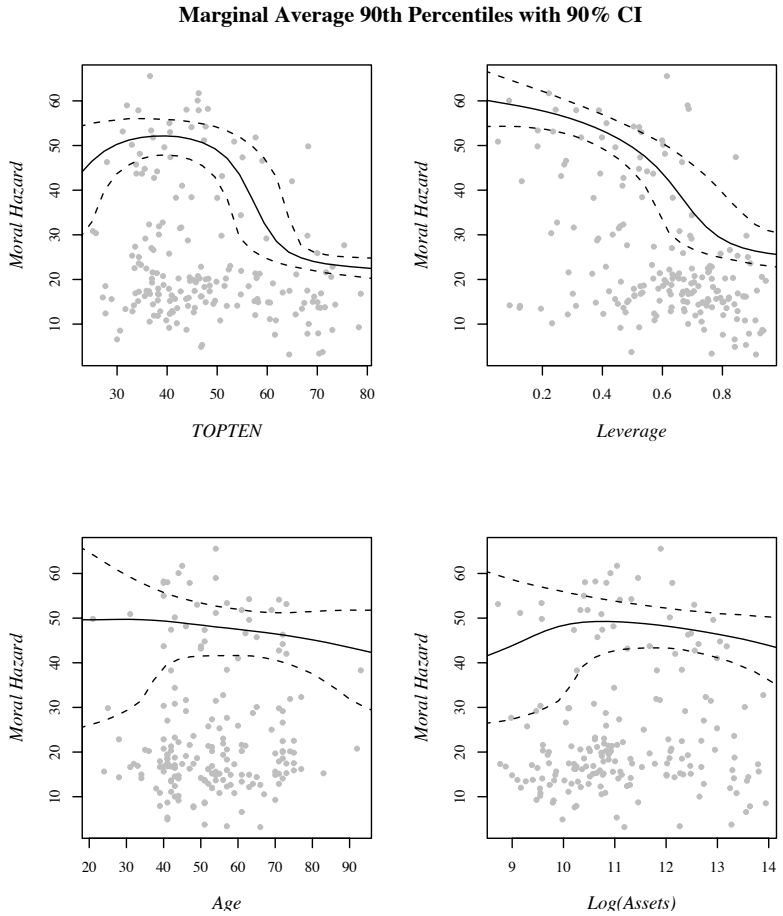


Figure 2.6. Posterior mean and 90% interval estimates for 90th percentile regression for *MH*5 conditional on each individual covariate. Data scatterplots are shown in grey.

Quantile regression: data example

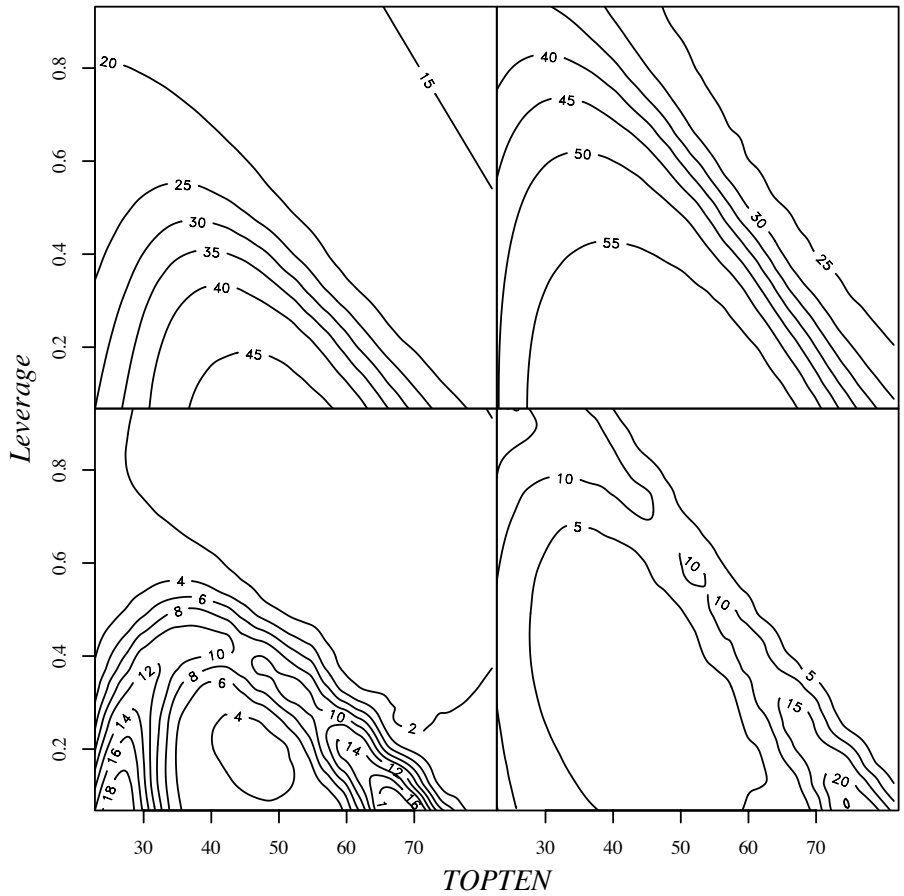


Figure 2.7. Posterior estimates of median surfaces (left column) and 90th percentile surfaces (right column) for *MH*5 conditional on *Leverage* and *TOPTEN*. The posterior mean is shown on the top row and the posterior interquartile range on the bottom.

Quantile regression: data example

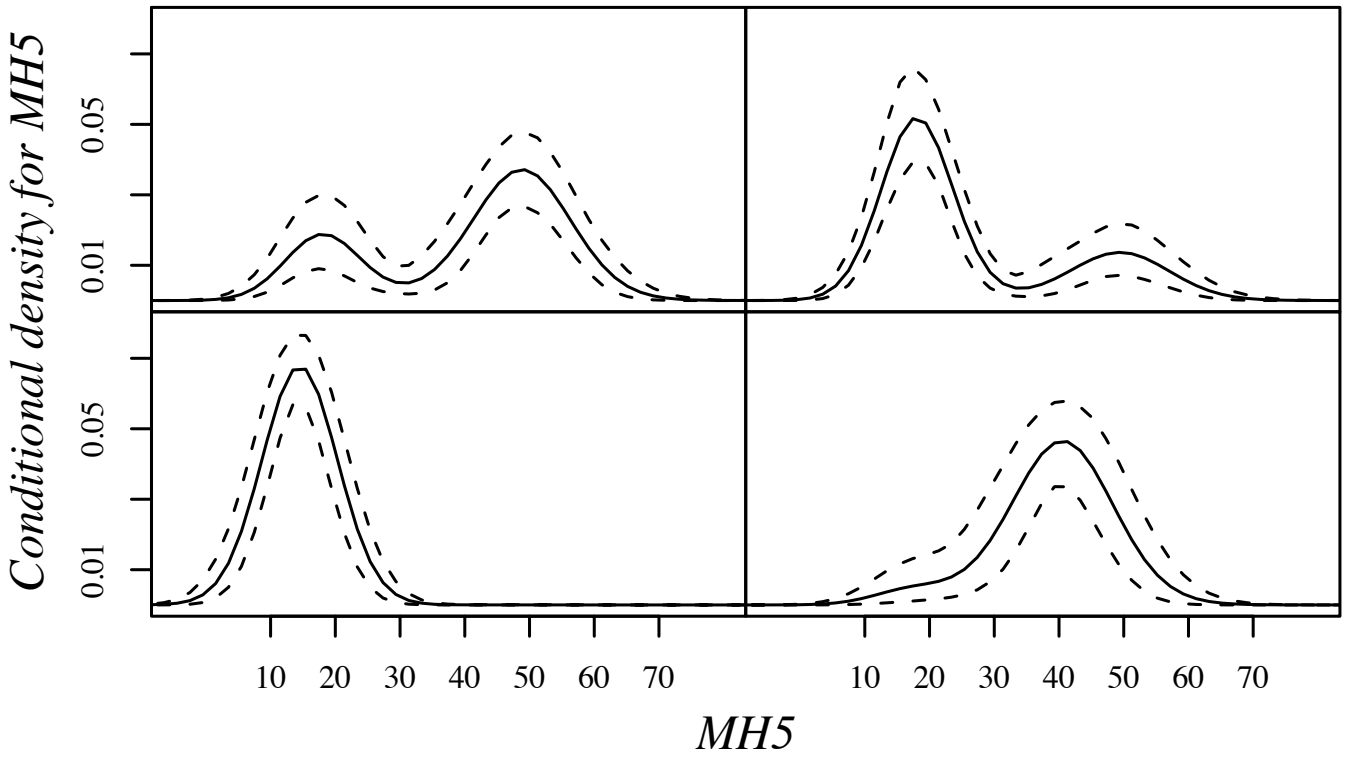


Figure 2.8. Posterior mean and 90% interval estimates for response densities $f(y \mid \boldsymbol{x}_0; G)$ conditional on four combinations of values $\boldsymbol{x}_0$ for the covariate vector (*TOPTEN*, *Leverage*, *Age*, log(*Assets*))

DP mixture density regression: applications

- Regression modeling with categorical responses (Shahbaba and Neal, 2009; Dunson and Bhattacharya, 2011; Hannah et al., 2011; DeYoreo and Kottas, 2015, 2018a,b).
- Functional data analysis through density estimation (Rodriguez et al., 2009).
- Markov switching regression (Taddy and Kottas, 2009), and fully nonparametric quantile regression (Taddy and Kottas, 2010).
- Product partition models with regression on covariates (Müller and Quintana, 2010; Park and Dunson, 2010), and regression modeling with *enriched* DP priors (Wade et al., 2014).
- Nonparametric survival regression (Poynor and Kottas, 2017).

Modeling for multivariate ordinal data

- Values of k ordinal categorical variables $Y_1, \ldots, Y_k$ recorded for n subjects:
  - $C_j \geq 2$: number of categories for the j-th variable, $j = 1, \ldots, k$.
  - $n_{\ell_1 \cdots \ell_k}$: number of observations such that

$$\mathbf{Y} = (Y_1, \ldots, Y_k) = (\ell_1, \ldots, \ell_k).$$

  - $p_{\ell_1 \cdots \ell_k} = \Pr(Y_1 = \ell_1, \ldots, Y_k = \ell_k)$ is the classification probability for the $(\ell_1, \ldots, \ell_k)$ cell.

- The data can be summarized in a k-dimensional contingency table with $C = \prod_{j=1}^{k} C_j$ cells, with frequencies $\{n_{\ell_1 \cdots \ell_k}\}$ constrained by $\sum_{\ell_1 \cdots \ell_k} n_{\ell_1 \cdots \ell_k} = n$.

Modeling for multivariate ordinal data

- A possible modeling strategy (alternative to log-linear models) involves the introduction of k continuous latent variables $\mathbf{Z} = (Z_1, \ldots, Z_k)$ whose joint distribution yields through discretization the classification probabilities for the table cells,

$$p_{\ell_1 \cdots \ell_k} = \Pr\left(\bigcap_{j=1}^{k} \left\{ \gamma_{j,\ell_j - 1} < Z_j \leq \gamma_{j,\ell_j} \right\} \right)$$

  for cutoff points $-\infty = \gamma_{j,0} < \gamma_{j,1} < \cdots < \gamma_{j,C_j - 1} < \gamma_{j,C_j} = \infty$, for each $j = 1, \ldots, k$ (e.g., Johnson and Albert, 1999).

- Common distributional assumption: $\mathbf{Z} \sim N_k(\mathbf{0}, \mathbf{S})$ (probit model).
  - $\rho_{st} = \text{Corr}(Z_s, Z_t) = 0$, $s \neq t$, implies independence of the corresponding categorical variables.
  - Coefficients ρ_{st}, $s \neq t$: *polychoric correlation coefficients* (traditionally used in the social sciences as a measure of association).

Modeling for multivariate ordinal data

- Richer modeling and inference based on normal DP mixtures for the latent variables $\mathbf{Z}_i$ associated with data vectors $\mathbf{Y}_i$, $i = 1, \ldots, n$.
- Model $\mathbf{Z}_i \mid G$ i.i.d. f, with $f(\cdot \mid G) = \int N_k(\cdot \mid \boldsymbol{m}, \boldsymbol{S}) \mathrm{d}G(\boldsymbol{m}, \boldsymbol{S})$, where

$$G \mid \alpha, \boldsymbol{\lambda}, \boldsymbol{\Sigma}, \boldsymbol{D} \sim \mathrm{DP}(\alpha, G_0(\boldsymbol{m}, \boldsymbol{S}) = \mathrm{N}_k(\boldsymbol{m} \mid \boldsymbol{\lambda}, \boldsymbol{\Sigma}) \mathrm{IW}_k(\boldsymbol{S} \mid \nu, \boldsymbol{D}))$$

- Advantages of the DP mixture modeling approach:
  - Can accommodate essentially any pattern in k-dimensional contingency tables.
  - Allows local dependence structure to vary accross the contingency table.
  - Implementation does not require random cutoffs (so the complex updating mechanisms for cutoffs are not needed).

Modeling for multivariate ordinal data: data example

- A data set on *Interrater Agreement*: data on the extent of scleral extension (extent to which a tumor has invaded the sclera or "white of the eye") as coded by two raters for each of $n = 885$ eyes.
- The coding scheme uses five categories: 1 for "none or innermost layers"; 2 for "within sclera, but does not extend to scleral surface"; 3 for "extends to scleral surface"; 4 for "extrascleral extension without transection"; and 5 for "extrascleral extension with presumed residual tumor in the orbit".
- Results under the DP mixture model (and, for comparison, using also a probit model).
- The $(0.25, 0.5, 0.75)$ posterior percentiles for n^* are $(6, 7, 8)$; in fact, $\Pr(n^* \geq 4 \mid \text{data}) = 1$.

Modeling for multivariate ordinal data: data example

For the interrater agreement data, observed cell relative frequencies (in bold) and posterior summaries for table cell probabilities (posterior mean and 95% central posterior intervals). Rows correspond to rater A and columns to rater B.

.3288 .3264 (.2940, .3586)	**.0836** .0872 (.0696, .1062)	**.0011** .0013 (.0002, .0041)	**.0011** .0020 (.0003, .0055)	**.0011** .0008 (.0, .0033)
.2102 .2136 (.1867, .2404)	**.2893** .2817 (.2524, .3112)	**.0079** 0.0080 (.0033, .0146)	**.0079** .0070 (.0022, .0143)	**.0034** .0030 (.0006, .0074)
.0023 .0021 (.0004, .0059)	**.0045** .0060 (.0021, .0118)	**.0** .0016 (.0004, .0037)	**.0023** .0023 (.0004, .0059)	**.0** .0009 (.0, .0030)
.0034 .0043 (.0012, .0094)	**.0113** .0101 (.0041, .0185)	**.0011** .0023 (.0004, .0058)	**.0158** .0142 (.0069, .0238)	**.0023** .0027 (.0006, .0066)
.0011 .0013 (.0001, .0044)	**.0079** .0071 (.0026, .0140)	**.0011** .0020 (.0003, .0054)	**.0090** .0084 (.0033, .0159)	**.0034** .0039 (.0011, .0090)

Modeling for multivariate ordinal data: data example

- Posterior predictive distributions $p(\mathbf{Z}_0 \mid \text{data})$ (see Figure 2.9) – DP mixture version is based on the posterior predictive distribution for corresponding mixing parameter $(\mathbf{m}_0, \mathbf{S}_0)$.
- Inference for the association between the ordinal variables:
  - For example, Figure 2.9 shows posteriors for ρ_0, the correlation coefficient implied in $\mathbf{S}_0$.
  - The probit model does not capture successfully the association of the ordinal variables, since it fails to recognize the clustering suggested by the data (revealed by the DP mixture model).
- Figure 2.10 shows inferences for log-odds ratios,

$$\psi_{ij} = log p_{i,j} + \log p_{i+1,j+1} - \log p_{i,j+1} - \log p_{i+1,j}.$$

- Utility of mixture modeling for this data example: one of the clusters dominates the others, but identifying the other three is important; one of them corresponds to agreement for large values in the coding scheme; the other two indicate regions of the table where the two raters tend to agree less strongly.

Modeling for multivariate ordinal data: data example

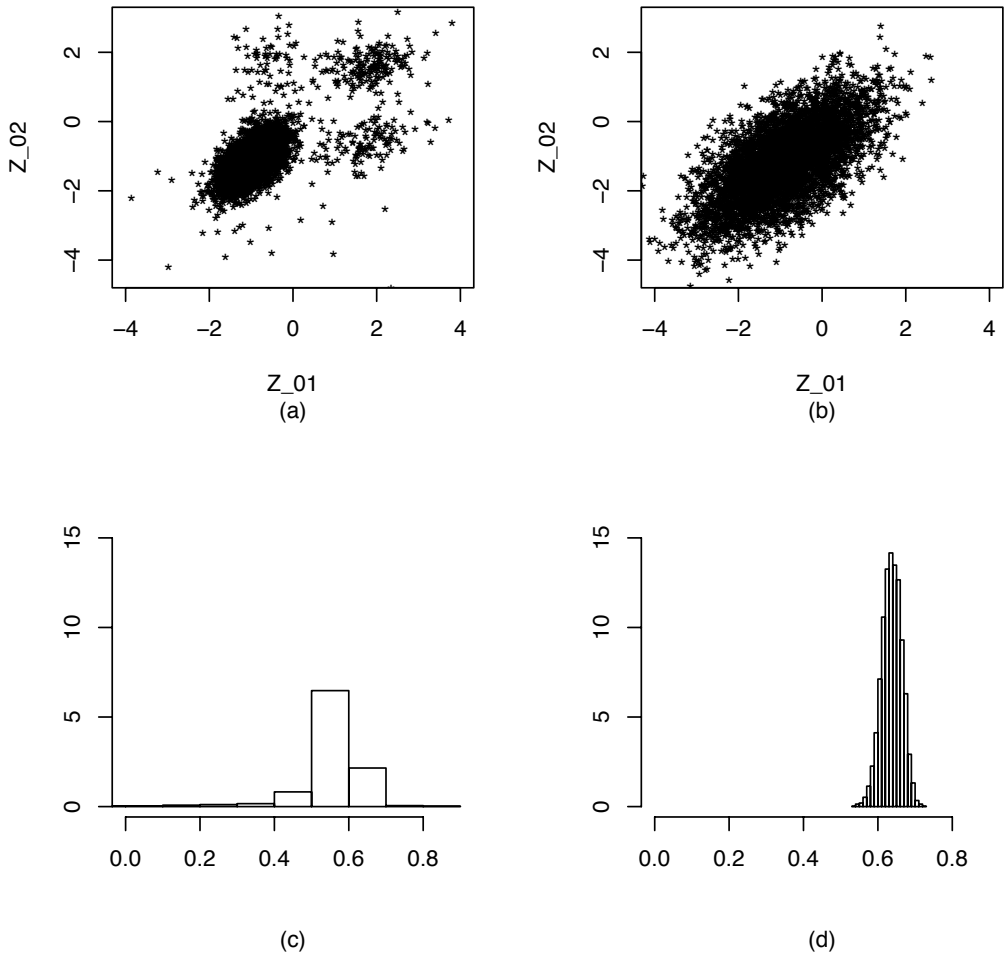


Figure 2.9. For the interrater agreement data, draws from $p(\mathbf{Z}_0 \mid \text{data})$ and $p(\rho_0 \mid \text{data})$ under the DP mixture model (panels (a) and (c), respectively) and the probit model (panels (b) and (d), respectively).

Modeling for multivariate ordinal data: data example

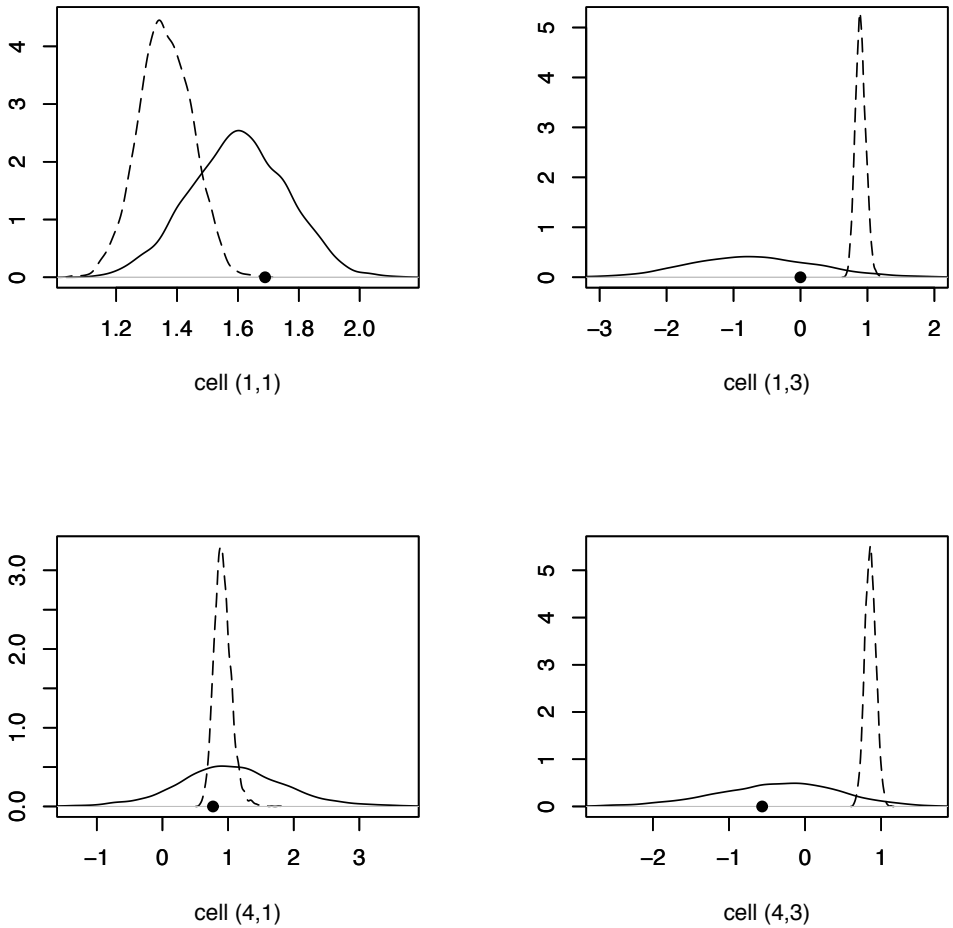


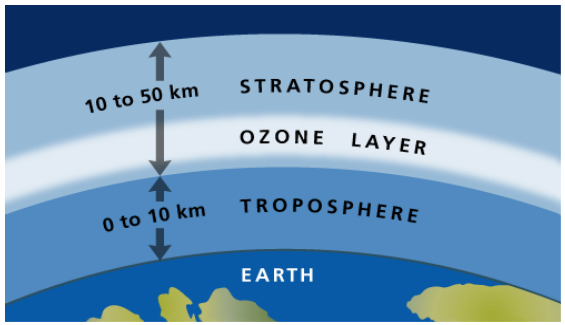
Figure 2.10. For the interrater agreement data, posteriors for four log-odds ratios under the DP mixture model (solid lines) and the probit model (dashed lines). The circles denote the corresponding empirical log-odds ratios.

Nonparametric multivariate ordinal regression

- k ordinal variables $\mathbf{Y} = (Y_1, \ldots, Y_k)$, with $y_j \in \{1, \ldots, C_j\}$, and p (continuous) covariates $\mathbf{X} = (X_1, \ldots, X_p)$.
- Again, $Y_j = \ell$ if-f $\gamma_{j,\ell-1} < Z_j \leq \gamma_{j,\ell}$, for $j = 1, ..., k$, and $\ell = 1, ..., C_j$.
- Now, model the joint distribution of the latent continuous responses, $\mathbf{Z} = (Z_1, \ldots, Z_k)$, and the covariates, $\mathbf{X}$, with a multivariate normal DP mixture $\rightarrow$ implies a regression model, $\Pr(\mathbf{Y} \mid \mathbf{x})$, which is a mixture of probit regressions with covariate-dependent weights.
- Large support under fixed cut-offs:
  - for any mixed ordinal-continuous distribution, $p_0(\boldsymbol{x}, \boldsymbol{y})$, that satisfies certain regularity conditions, the prior model assigns positive probability to all Kullback-Leibler (KL) neighborhoods of $p_0(\boldsymbol{x}, \boldsymbol{y})$, as well as to all KL neighborhoods of the implied conditional distribution, $p_0(\boldsymbol{y} \mid \boldsymbol{x})$.

Ozone concentration data example

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



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

Ozone concentration data example

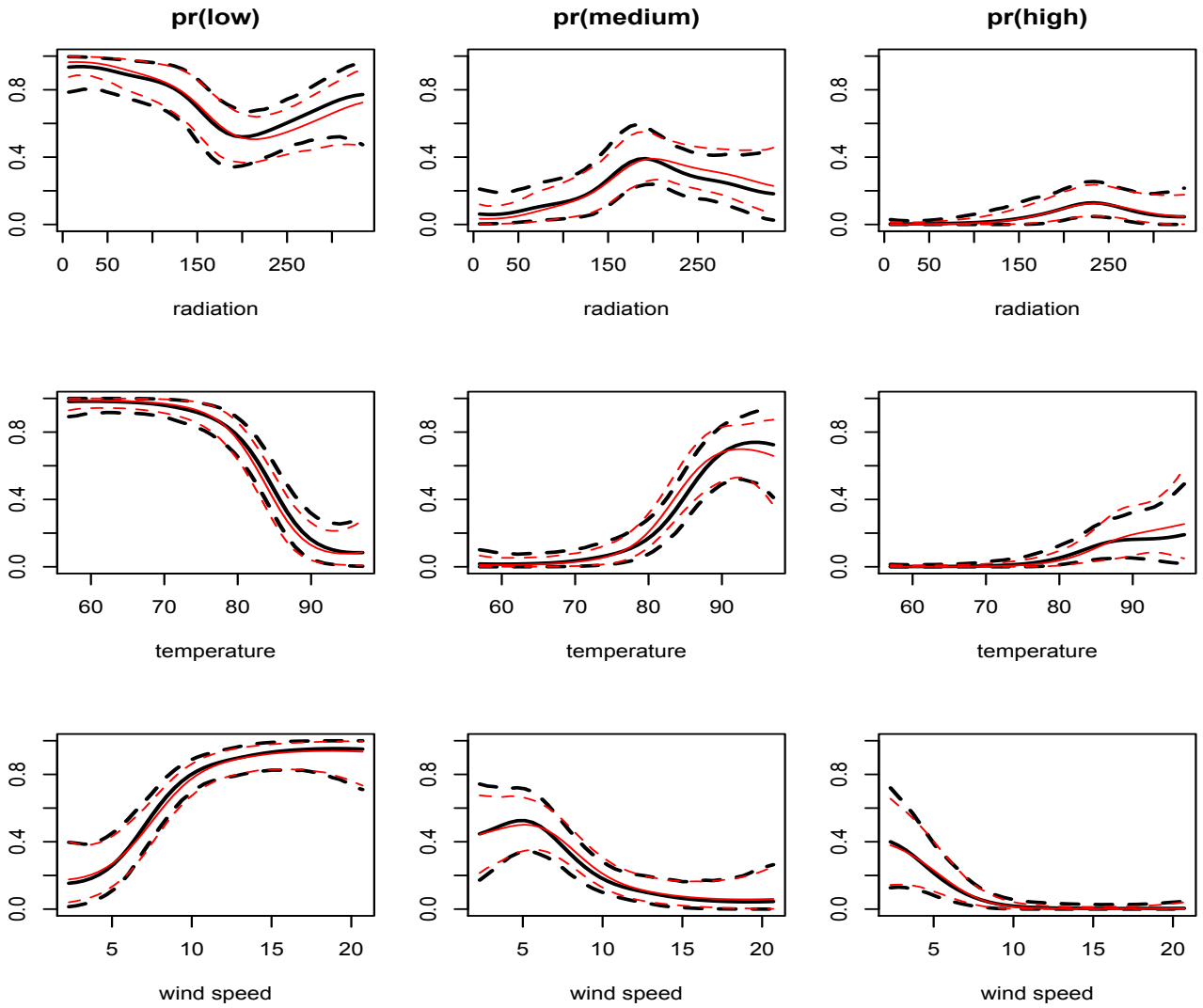


Figure 2.11. Posterior mean (solid) and 95% interval estimates (dashed) for $\Pr(Y = \ell \mid x_m, G)$ (black) compared to $\Pr(\gamma_{\ell-1} < Z \leq \gamma_\ell \mid x_m, G)$ (red).

Ozone concentration data example

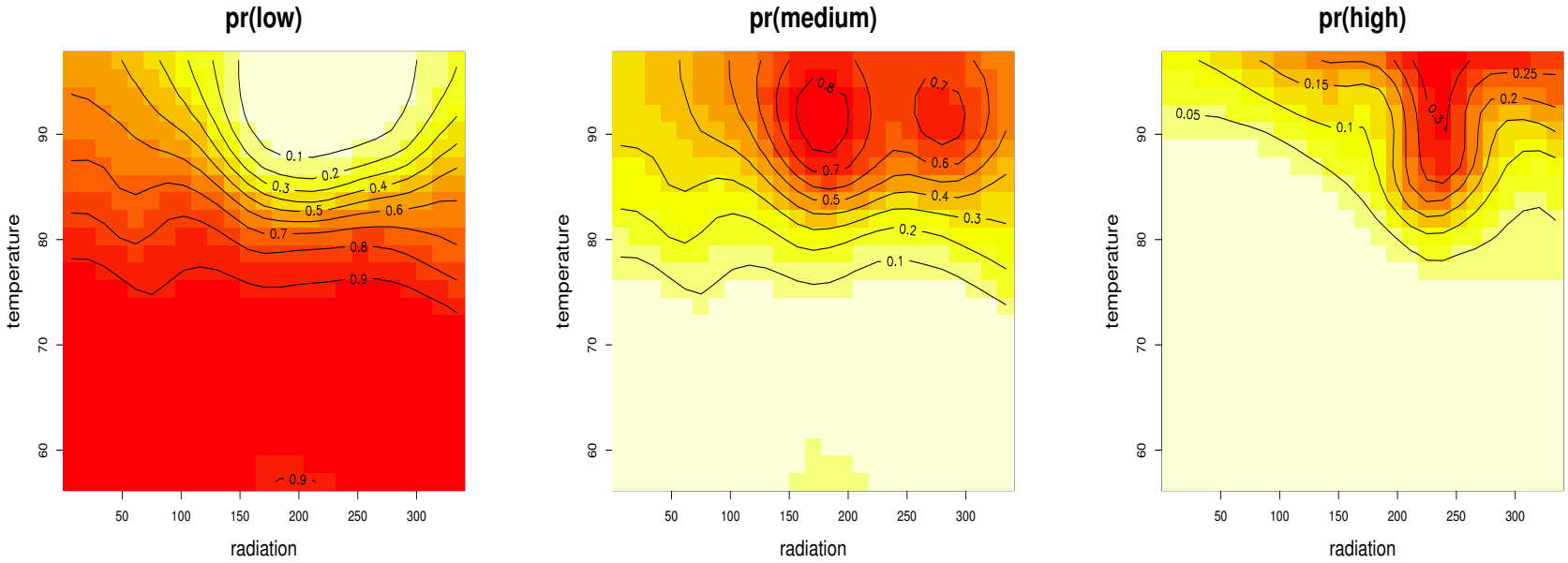


Figure 2.12. Posterior mean estimates for $\Pr(Y = \ell \mid x_1, x_2, G)$, for $\ell = 1, 2, 3$, corresponding to low (left), medium (middle) and high (right). Red represents a value of 1, white represents 0.

Nonparametric inference for Poisson processes

- Point processes are stochastic process models for events that occur separated in time or space.
- Applications of point process modeling in traffic engineering, software reliability, neurophysiology, weather modeling, forestry, ...
- Poisson processes, along with their extensions (Poisson cluster processes, marked Poisson processes, etc.), play an important role in the theory and applications of point processes. (e.g., Kingman, 1993; Guttorp, 1995; Moller & Waagepetersen, 2004).
- Bayesian nonparametric work based on gamma processes, weighted gamma processes, and Lévy processes (e.g., Lo & Weng, 1989; Kuo & Ghosh, 1997; Wolpert & Ickstadt, 1998; Gutiérrez-Peña & Nieto-Barajas, 2003; Ishwaran & James, 2004).

Definition of Poisson processes on the real line

- For a point process over time, let $N(t)$ be the number of event occurrences in the time interval $(0, t]$.
- The point process $\mathcal{N} = \{N(t) : t \geq 0\}$ is a non-homogeneous Poisson process (NHPP) if:
  - For any $t > s \geq 0$, $N(t) - N(s)$ follows a Poisson distribution with mean $\Lambda(t) - \Lambda(s)$.
  - $\mathcal{N}$ has independent increments, i.e., for any $0 \leq t_1 < t_2 \leq t_3 < t_4$, $N(t_2) - N(t_1)$ and $N(t_4) - N(t_3)$ are independent random variables.
- Λ is the mean measure (or cumulative intensity function) of the NHPP.
- For any $t \in R^+$, $\Lambda(t) = \int_0^t \lambda(u)\mathrm{d}u$, where λ is the NHPP intensity function – λ is a non-negative and locally integrable function (i.e., $\int_B \lambda(u)\mathrm{d}u < \infty$, for all bounded $B \subset \mathbb{R}^+$).
- So, from a modeling perspective, the main functional of interest for a NHPP is its intensity function.

Nonparametric inference for Poisson processes

- Consider a NHPP observed over the time interval $(0, T]$ with events that occur at times $0 < t_1 < t_2 < \ldots < t_n \leq T$.
- The likelihood for the NHPP intensity function λ is proportional to

$$\exp\left\{-\int_0^T \lambda(u)\mathrm{d}u\right\}\prod_{i=1}^{n}\lambda(t_i).$$

- **Key observation:** $f(t) = \lambda(t)/\gamma$, where $\gamma = \int_0^T \lambda(u)\mathrm{d}u$, is a density function on $(0, T)$.
- Hence, a nonparametric prior model for f, with a parametric prior for γ, will induce a semiparametric prior for λ.
- Since γ only scales λ, it is f that determines the shape of the intensity function λ.

Nonparametric inference for Poisson processes

- **Beta DP mixture model** for f:

$$f(t) \equiv f(t \mid G) = \int \text{Beta}(t \mid \mu, \tau) \text{d}G(\mu, \tau), \quad G \sim \text{DP}(\alpha, G_0)$$

  where $\text{Beta}(t \mid \mu, \tau)$ is the Beta density on $(0, T)$ with mean $\mu \in (0, T)$ and scale parameter $\tau > 0$, and $G_0(\mu, \tau) = \text{Uni}(\mu \mid 0, T)$ $\text{IG}(\tau \mid c, \beta)$ with random scale parameter β.

- Flexible density shapes through mixing of Betas (e.g., Diaconis and Ylvisaker, 1985) – Beta mixture model avoids edge effects (a drawback of the normal DP mixture model in this setting).

- Full Bayesian model:

$$e^{-\gamma}\gamma^n \left\{ \prod_{i=1}^{n} \int \text{Beta}(t_i \mid \mu_i, \tau_i) \text{d}G(\mu_i, \tau_i) \right\} p(\gamma) \text{DP}(G \mid \alpha, G_0(\beta)) p(\alpha) p(\beta)$$

- Reference prior for γ, $p(\gamma) \propto \gamma^{-1}$.

Nonparametric inference for Poisson processes

- Letting $\boldsymbol{\theta} = \{(\mu_i, \tau_i) : i = 1, \ldots, n\}$, we have

$$p(\gamma, G, \boldsymbol{\theta}, \alpha, \beta \mid \text{data}) = p(\gamma \mid \text{data}) p(G \mid \boldsymbol{\theta}, \alpha, \beta) p(\boldsymbol{\theta}, \alpha, \beta \mid \text{data})$$

  where:
  - $p(\gamma \mid \text{data})$ is a gamma$(n, 1)$ distribution.
  - MCMC is used to sample from $p(\boldsymbol{\theta}, \alpha, \beta \mid \text{data})$.
  - $p(G \mid \boldsymbol{\theta}, \alpha, \beta)$ is a DP with updated parameters (can be sampled as discussed earlier).

- Full posterior inference for λ, Λ, and any other NHPP functional.

- Extensions to inference for spatial NHPP intensities, using DP mixtures with bivariate Beta kernels (Kottas and Sansó, 2007).

Data examples

- Example for temporal NHPPs: times of 191 explosions in mines, leading to coal-mining disasters with 10 or more men killed, over a time period of 40,550 days, from 15 March 1851 to 22 March 1962.
- Specification for $\text{DP}(\alpha, G_0(\mu, \tau \mid \beta) = \text{Uni}(\mu \mid 0, T)\text{IG}(\tau \mid 2, \beta))$.
  - gamma(a_α, b_α) prior for α.
  - Exponential prior for β – its mean can be specified using a prior guess at the range, R, of the event times t_i (e.g., $R = T$ is a possible default choice).
- Inference for the NHPP intensity under three prior choices: priors for β and α based on $R = T$, $\text{E}(n^*) \approx 7$; $R = T$, $\text{E}(n^*) \approx 15$; and $R = 1.5T$, $\text{E}(n^*) \approx 7$.
- Examples for spatial NHPPs, using two forestry data sets:
  - locations of 62 redwood seedlings in a square of 23 m;
  - locations of 514 maple trees in a 19.6 acre square plot in Lansing Woods, Clinton County, MI.

Data examples

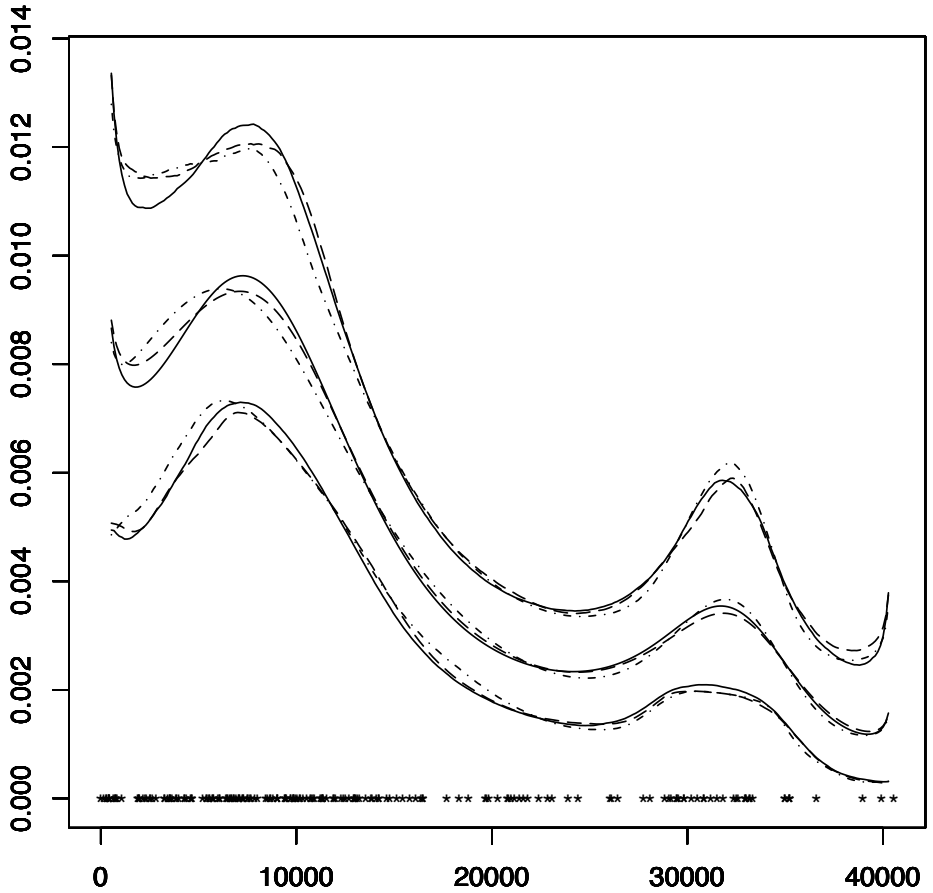


Figure 2.13. Coal-mining disasters data. Posterior point and 95% interval estimates for the intensity function under three prior settings. The observed times of the 191 explosions in mines are plotted on the horizontal axis.

Data examples

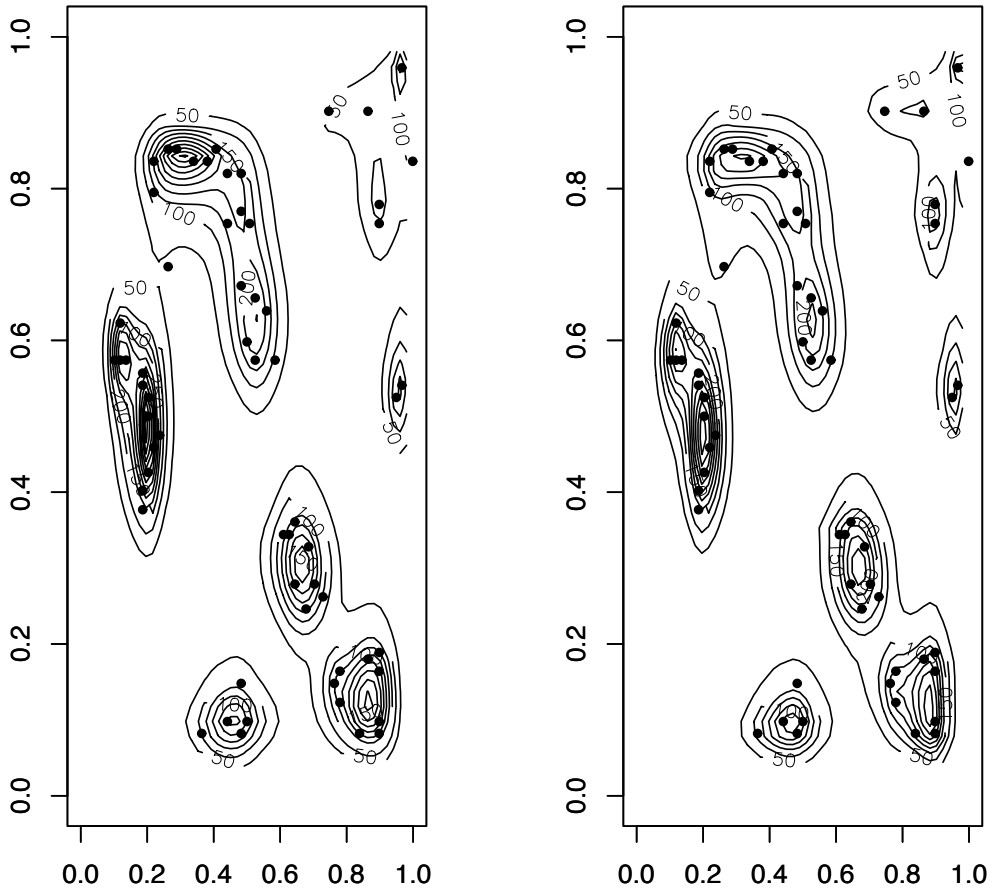


Figure 2.14. Redwood seedlings data. Contour plots of posterior mean intensity estimates under two different priors for α. The dots indicate the locations of the redwood seedlings.

Data examples

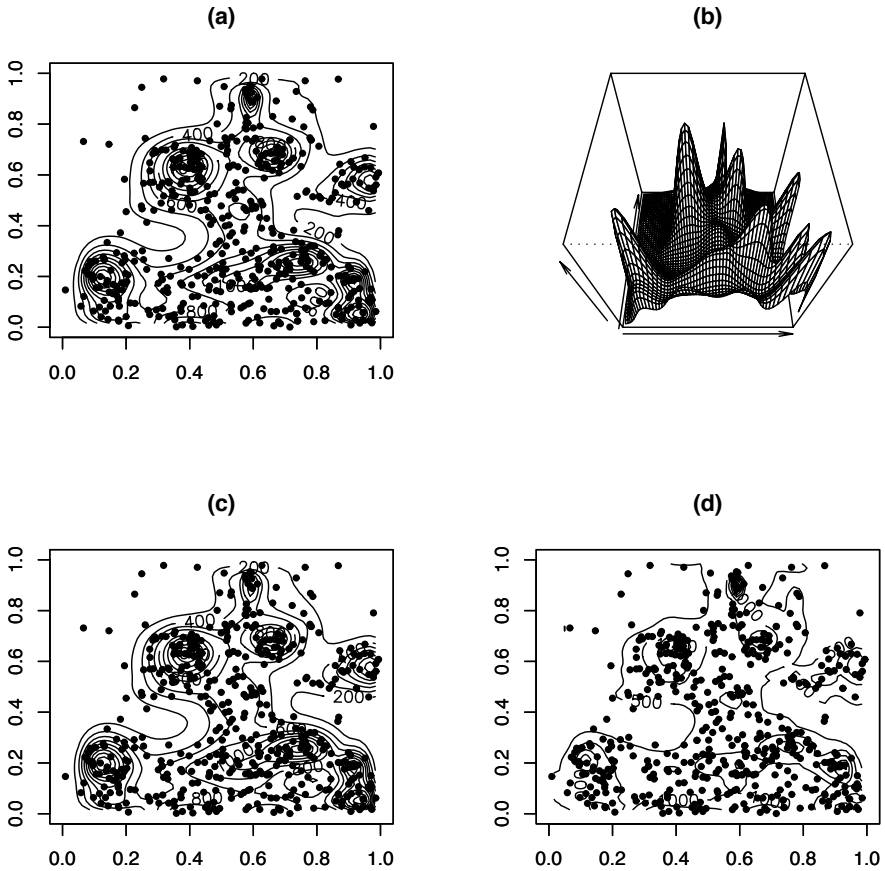


Figure 2.15. Maples data. Panels (a) and (b) include the posterior mean intensity estimate (contour and perspective plot, respectively). Panels (c) and (d) show contour plots for the posterior median and interquartile range intensity estimates, respectively. The dots denote the locations of the maple trees.

Nonparametric modeling for NHPPs: further work

- Applications to neuronal data analysis (Kottas and Behseta, 2010; Kottas et al., 2012).
- Inference for marked Poisson processes (Taddy & Kottas, 2012).
- Dynamic modeling for spatial NHPPs (Taddy, 2010).
- Risk assessment of extremes from spatially dependent environmental time series (Kottas et al., 2012) and from correlated financial markets (Rodriguez et al., 2017).
- Dynamic modeling for time-varying seasonal intensities, with an application to predicting hurricane damage (Xiao et al., 2015).