Bayesian nonparametrics

Dr. Jarad Niemi

STAT 615 - Iowa State University

November 30, 2017

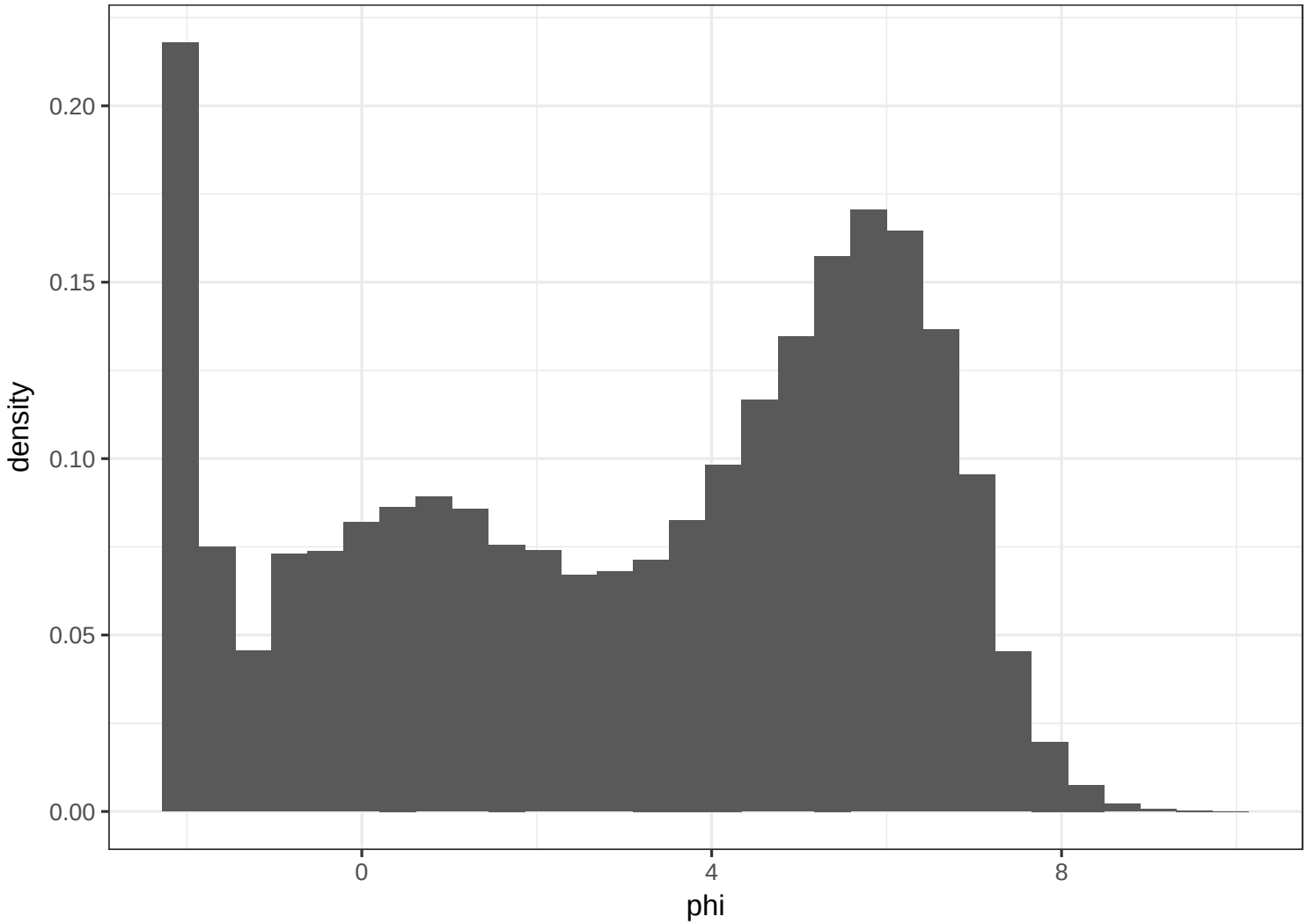
Bayesian nonparametrics

There are two main approaches to Bayesian nonparametrics for density estimation

- Dirichlet process and
- Polya trees

See Muller and Mitra (2013) for a general overview of all Bayesian nonparametric problems, e.g. density estimation, clustering, regression, random effects distributions, etc.

Motivation



Goal

Let Y_i come from an unknown probability measure $\mathcal{G}$, i.e. $Y_i \sim \mathcal{G}$. As a Bayesian, the natural approach is to put a prior on $\mathcal{G}$. That is, we want to make statements like

$$P(Y_i \in A) = \mathcal{G}(A)$$

for any set A.

Dirichlet process

One approach is to use a Dirichlet process (Ferguson 1973). We write

$$\mathcal{G} \sim DP(aG_0)$$

where

- $a > 0$ is concentration (or total mass) parameter and
- G_0 is the base measure, i.e. a probability distribution defined on the support of $\mathcal{G}$.

For any partition $A_1, \ldots, A_K$ of the sample space S, the probability vector $[\mathcal{G}(A_1), \ldots, \mathcal{G}(A_K)]$ follows a Dirichlet distribution, i.e.

$$[\mathcal{G}(A_1), \ldots, \mathcal{G}(A_K)] \sim Dir([aG_0(A_1), \ldots, aG_0(a_K)]).$$

Thus

- $E[\mathcal{G}(A_1)] = G_0(A_1)$ and
- $Var[\mathcal{G}(A_1)] = \frac{G_0(A_1)[1-G_0(A_1)]}{1+a}$.

Conjugacy of the Dirichlet process

Assume

$$Y_i \overset{ind}{\sim} \mathcal{G} \qquad \text{and} \qquad \mathcal{G} \sim DP(aG_0)$$

then for any partition $\{A_1, \ldots, A_K\}$, we have

$$[\mathcal{G}(A_1), \ldots, \mathcal{G}(A_K)]|y \sim Dir\left([aG_0(A_1) + \textstyle\sum_{i=1}^n \mathrm{I}(y_i \in A_1), \ldots, aG_0(A_K) + \sum_{i=1}^n \mathrm{I}(y_i \in A_K)]\right)$$

and thus

$$\mathcal{G}|y \sim DP\left(aG_0 + \sum_{i=1}^n \delta_{y_i}\right)$$

which has

$$E[\mathcal{G}(A)|y] = \left(\frac{a}{a+n}\right) G_0(A) + \left(\frac{n}{a+n}\right) \sum_{i=1}^n \frac{1}{n} \mathrm{I}(y_i \in A)$$

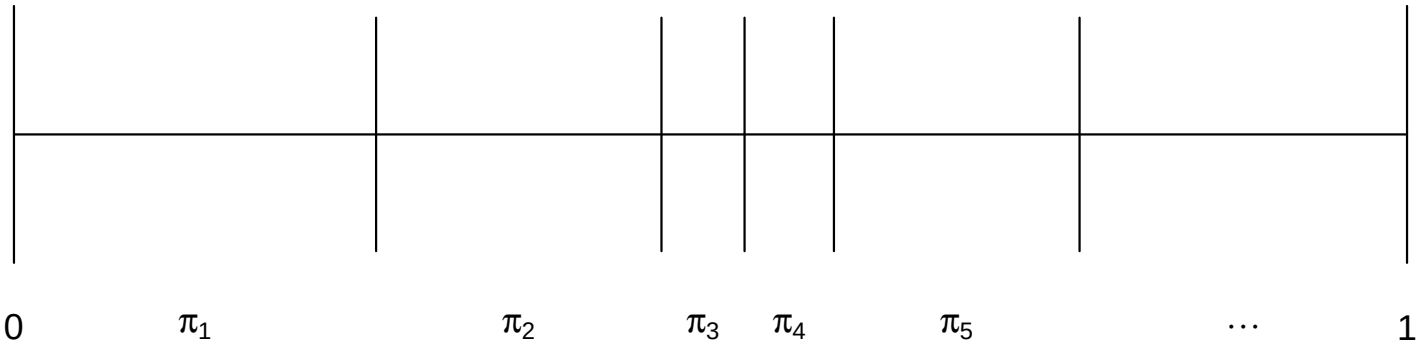
Stick-breaking representation

A constructive representation of the Dirichlet process is the stick-breaking representation. Assume $\mathcal{G} \sim DP(aG_0)$, then

$$\mathcal{G}(\cdot) = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h}(\cdot)$$

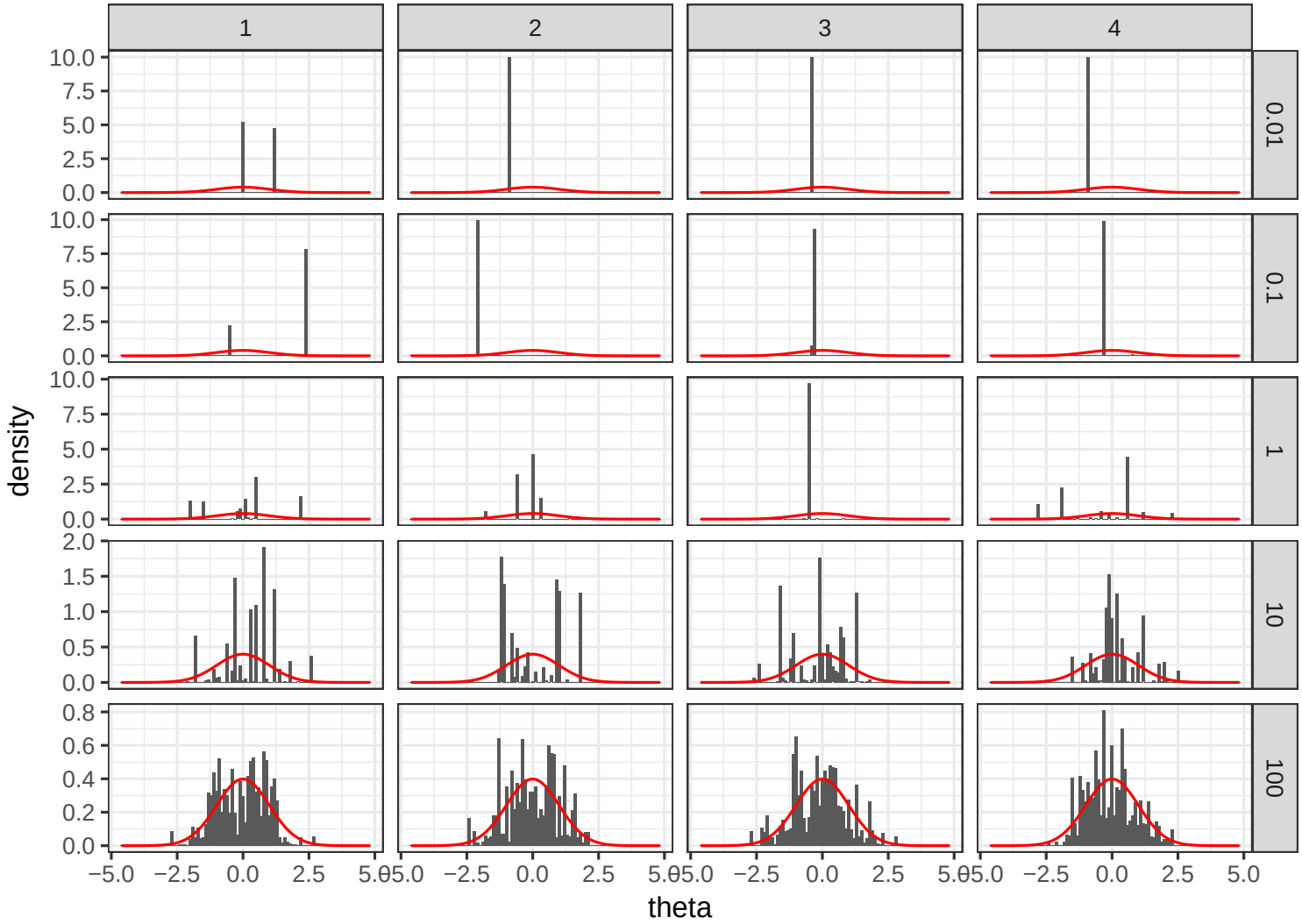
where $\pi \sim \text{stick}(a)$ and $\theta_h \overset{ind}{\sim} G_0$. The stick distribution is the following:

- $\pi_h = \nu_h \prod_{\ell<h}(1-\nu_\ell)$ and
- $\nu_h \overset{ind}{\sim} Be(1, a)$.



Realizations from a DP

Base measure is a standard normal. Realizations are across the columns and values for a are down the rows.



DP mixture

If we have an absolutely continuous distribution we are trying to approximate, then a DP is not reasonable. Thus, we may want to use a DP mixture, i.e.

$$Y_i \overset{ind}{\sim} p(\cdot|\theta_i), \quad \theta_i \overset{ind}{\sim} \mathcal{G}, \quad \mathcal{G} \sim DP(aG_0)$$

for some parametric model $p(\cdot|\theta)$.
Alternatively, if we use the stick-breaking construction, we have

$$Y_i \overset{ind}{\sim} p(\cdot|\theta_i), \qquad \theta_i \overset{ind}{\sim} \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h^*}$$

where $\theta_h^* \overset{ind}{\sim} G_0$ and $\pi \sim \text{stick}(a)$.

Finite approximation to the stick-breaking representation

For some $\epsilon > 0$, there exists an H such that $\sum_{h=H}^{\infty} \pi_h < \epsilon$ and components H and beyond can reasonably be ignored. The resulting model is

$$Y_i \overset{ind}{\sim} p(\cdot|\theta_i), \quad \theta_i \overset{ind}{\sim} \sum_{h=1}^{H} \pi_h \delta_{\theta_h^*}$$

where

- $\pi_h = V_h \prod_{\ell<h}(1 - V_\ell)$
- $V_h \overset{ind}{\sim} Be(1, a)$ for $h < H$, and
- $V_H = 1$.

Normal example

A DP mixture model for the marginal distribution for $Y_i = \phi_i$ is

$$Y_i \overset{ind}{\sim} N(\mu_i, \sigma_i^2) \quad \begin{pmatrix} \mu_i \\ \sigma_i^2 \end{pmatrix} \sim \sum_{h=1}^{H} \pi_h \delta_{(\mu_h, \sigma_h^{2*})}$$

where $\sum_{h=1}^{H} \pi_h = 1$.

Alternatively, we can introduce a latent variable $\zeta_i = h$ if observation i came from group h. Then

$$\begin{aligned} Y_i|\zeta_i = h &\overset{ind}{\sim} N(\mu_h, \sigma_h^2) \\ \zeta_i &\overset{ind}{\sim} Cat(H, \pi) \end{aligned}$$

where $\zeta \sim Cat(H, \pi)$ is a categorical random variable with $P(\zeta = h) = \pi_h$ for $h = 1, \ldots, H$ and $\pi = (\pi_1, \ldots, \pi_H)$.

Normal example

Let

$$Y_i \overset{ind}{\sim} N(\mu_i, \sigma_i^2), \qquad (\mu_i, \sigma_i^2) \overset{ind}{\sim} \sum_{h=1}^{H} \pi_h \delta_{(\mu_h^*, \sigma_h^{2*})}$$

where the base measure G_0 is

$$\mu_h^* | \sigma_h^{2*} \overset{ind}{\sim} N(m_h, v_h^2 \sigma_h^{2*}) \qquad \text{and} \qquad \sigma_h^{2*} \overset{ind}{\sim} IG(c_h, d_h).$$

But since each (μ_i, σ_i^2) must equal (μ_h^*, σ_h^{2*}) for some h, we can rewrite the model as

$$Y_i \overset{ind}{\sim} \sum_{h=1}^{H} \pi_h N(\mu_h^*, \sigma_h^{2*})$$

with a prior that is equal to the base measure. Thus this model is equivalent to our finite mixture with the exception of the prior for π.

MCMC - Blocked Gibbs sampler

The steps of a Gibbs sampler with stationary distribution

$$p(\pi, \mu, \sigma^2, \zeta|y) \propto p(y|\zeta, \mu, \sigma^2)p(\zeta|\pi)p(\mu|\sigma^2)p(\sigma^2)p(\pi)$$

has steps

1. For $i = 1, \ldots, n$, independently sample ζ_i from its full conditional

$$P(\zeta_i = h| \ldots) \propto \pi_h N(y_i; \mu_h^*, \sigma_h^{2*})$$

2. Jointly sample π and μ, σ^2 because they are conditionally independent.
   a. Sample $V_h \overset{ind}{\sim} Be(1 + Z_h, a + Z_h^+)$ for $V = 1, \ldots, H-1$ where $Z_h = \sum_{i=1}^n \mathrm{I}(\zeta_i = h)$ and $Z_h^+ = \sum_{h'=h+1}^H Z_{h'}$ and set $V_H = 1$. Then calculate $\pi_h = V_h \prod_{\ell<h}(1 - V_\ell)$.
   b. For $h = 1, \ldots, H$, sample μ_h, σ_h^2 from their full conditional

$$\mu_h^*|\sigma_h^{2*} \overset{ind}{\sim} N(m_h', v_h'^2) \quad \sigma_h^{2*} \overset{ind}{\sim} IG(c_h', d_h')$$

   where $m_h', v_h'^2, c_h'$, and d_h' are exactly the same as in the normal finite mixture MCMC.

```
library("rjags")
dp_normal_blocked = "
model {
  for (i in 1:n) {
    y[i] ~ dnorm(mu[zeta[i]], tau[zeta[i]])
    zeta[i] ~ dcat(pi[])
  }

  for (h in 1:H) {
    mu[h] ~ dnorm(2,1/3)
    tau[h] ~ dgamma(.1,.1)
    sigma[h] <- 1/sqrt(tau[h])
  }

  # Stick breaking
  for (h in 1:(H-1)) { V[h] ~ dbeta(1,a) }
  V[H] <- 1
  pi[1] <- V[1]
  for (h in 2:H) {
    pi[h] <- V[h] * (1-V[h-1]) * pi[h-1] / V[h-1]
  }
}"
```

```
tmp = hat[sample(nrow(hat), 1000),]
dat = list(n=nrow(tmp), H=25, y=tmp$phi, a=1)
```

```
jm = jags.model(textConnection(dp_normal_blocked), data = dat, n.chains = 3)
r = jags.samples(jm, c('mu','sigma','pi','zeta'), 1e3)
```

Monitor convergence of density

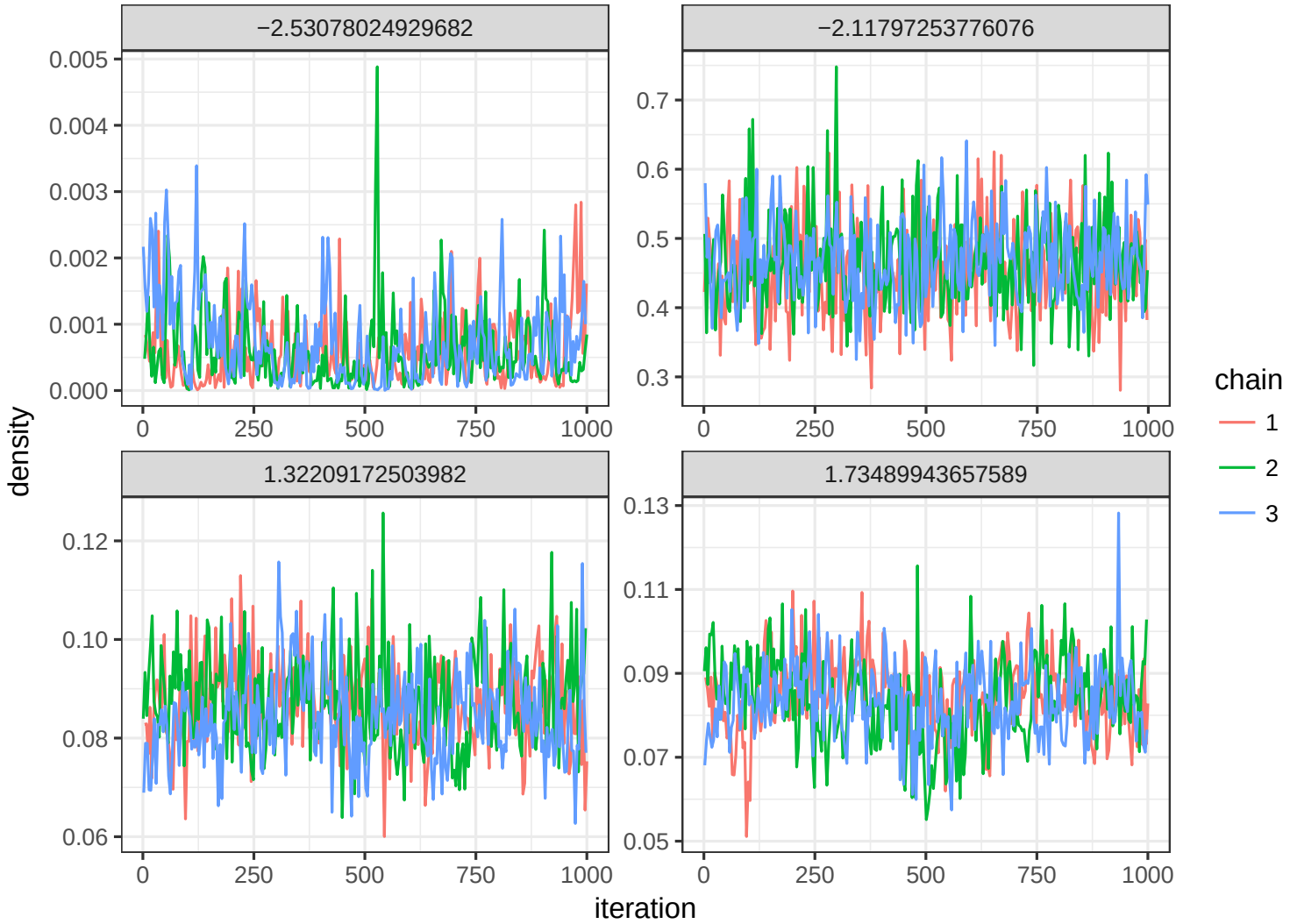
As previously discussed, the model as constructed as identifiability problems among the π_h, μ_h^*, and σ_h^{2*} due to label switching. What is identified in the model is the value of the density at any particular value.

So rather than directly monitoring the parameters, we will monitor the estimated density, i.e. at iteration m of the MCMC, the estimated density at location x is

$$\sum_{h=1}^{H} \pi_h^{(m)} N(x; \mu_h^{*(m)}, \sigma_h^{2*(m)}).$$

Monitoring this quantity at a variety of locations x will provide appropriate convergence assessment.

Monitor convergence of density



Monitoring the number of utilized components

Since we are using a finite approximation to the DP, we should monitor the index of the maximum occupied component (or the number of occupied clusters). If the finite approximation is reasonable, then this number will be smaller than H. If not, then H should be increased.

Specifically, at iteration m, we monitor

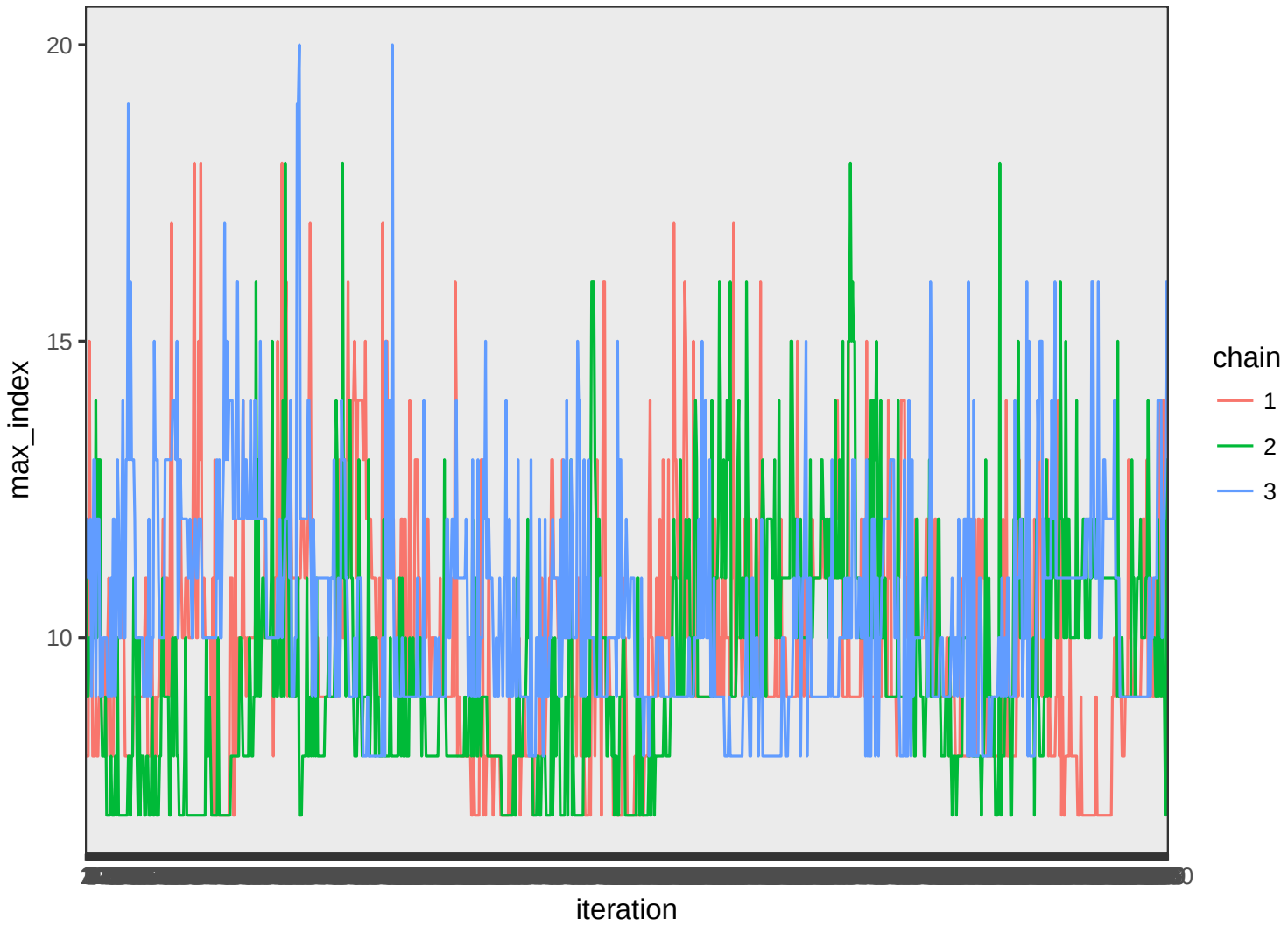
$$\max\{\zeta_1^{(m)}, \ldots, \zeta_n^{(m)}, \}$$

the index of the maximum occupied cluster, or

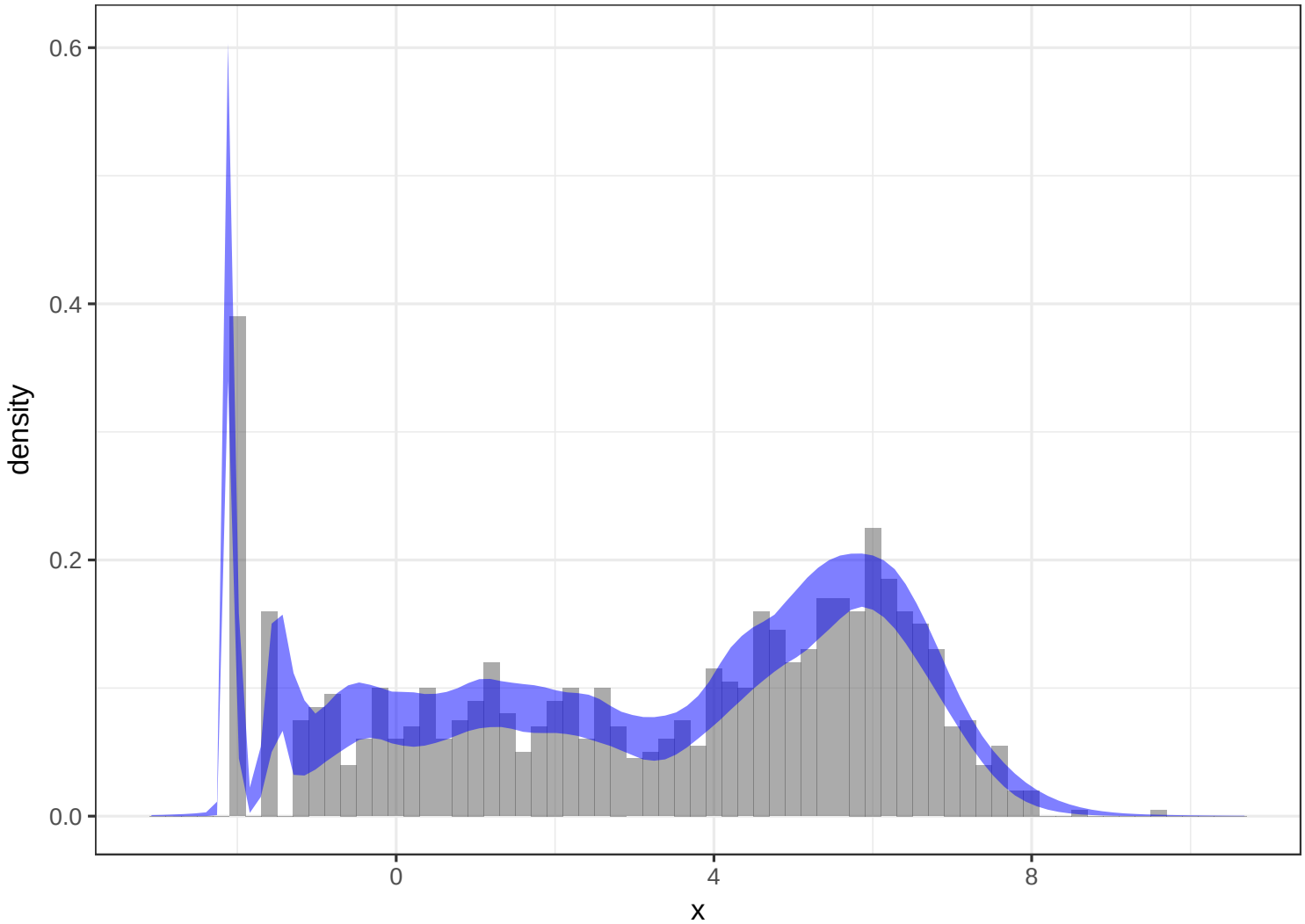
$$\sum_{h=1}^{H} \mathrm{I}(Z_h > 0),$$

the number of occupied clusters.

Monitoring the number of utilized components



Posterior density estimation



Chinese restaurant process

Rather than utilizing the finite approximation to the DP, we can use the DP directly, by marginalizing out $\mathcal{G}$. This results in a prior directly on $\theta_1, \ldots, \theta_n$ via

$$\theta_i|\theta_1, \ldots, \theta_{i-1} \sim \left(\frac{a}{a+i-1}\right) G_0(\theta_i) + \sum_{j=1}^{i-1} \left(\frac{1}{a+i-1}\right) \delta_{\theta_j}$$

The conditional prior for θ_i is

$$\theta_i|\theta_{-i} \sim \left(\frac{a}{a+n-1}\right) G_0(\theta_i) + \sum_{j\neq i} \left(\frac{1}{a+n-1}\right) \delta_{\theta_j}$$

or, equivalently,

$$\theta_i|\theta_{-i} \sim \left(\frac{a}{a+n-1}\right) G_0(\theta_i) + \sum_{h=1}^{H^{(-i)}} \left(\frac{n_h^{(-i)}}{a+n-1}\right) \delta_{\theta_h^*}$$

where $H^{(-i)}$ is the number of components without i and $n_h^{(-i)}$ is the number of observations in each component without i.

Marginalized Gibbs sampler

Using this Chinese restaurant process, we have the following $n+1$-step MCMC

1. For $i = 1, \ldots, n$, sample ζ_i from its full conditional

$$P(\zeta_i = h | \zeta_{-i}, \ldots) \propto \begin{cases} n_h^{(-i)} p(y_i | \theta_h^*) & h = 1, \ldots, H^{(-i)} \\ a \int p(y_i; \theta) dG_0(\theta) & h = H^{(-i)} + 1 \end{cases}$$

   If $\zeta_i = H^{(-i)} + 1$, then sample $\theta^*_{\zeta_i}$ from its posterior using y_i as the only observation.

2. For $h = 1, \ldots, H$, sample θ_h^* from their full conditional

$$\theta_h^* | \ldots \propto G_0(\theta_h^*) \prod_{i:\zeta_i = h} p(y_i | \theta_h^*)$$

   i.e. sample the parameters from their posteriors using only the data in that group.

Marginalized Gibbs sampler - Normal example

For the normal example, we have this $n+1$-step sampler

1. For $i = 1, \ldots, n$, sample ζ_i from its full conditional

$$P(\zeta_i = h | \zeta_{-i}, \ldots) \propto \begin{cases} n_h^{(-i)} N(y_i; \mu_h^*, \sigma_h^{2*}) & h = 1, \ldots, H^{(-i)} \\ a\, t_{2c}(y_i; m, v^2[d/c]) & h = H^{(-i)} + 1 \end{cases}$$

   If $\zeta_i = H^{(-i)} + 1$, then sample $\mu_{\zeta_i}^*, \sigma_{\zeta_i}^{2*}$ from its normal-inverse-gamma posterior using y_i as a the only observation.

2. For $h = 1, \ldots, H$, sample μ_h, σ_h^2 from their full conditional

$$\mu_h^* | \sigma_h^{2*} \overset{ind}{\sim} N(m_h', v_h'^2) \quad \sigma_h^{2*} \overset{ind}{\sim} IG(c_h', d_h')$$

   where $m_h', v_h'^2, c_h'$, and d_h' are exactly the same as in the normal finite mixture MCMC.

Putting a prior on the concentration parameter

If $\mathcal{G} \sim DP(aG_0)$, then the concentration parameter (a) controls the prior on the number of clusters. For example, if $a = 1$, then in the prior two randomly selected observations have a 0.5 probability of belonging to the same cluster. As a increases, then you have more clusters and more concentration around G_0. As a decreases, then you have fewer clusters and the data are more informative.

Rather than setting the concentration parameter, we can learn it. Let $\mathcal{G} \sim DP(\alpha G_0)$ and

$$\alpha \sim Ga(a, b)$$

then the full conditional for α is

$$\alpha | \ldots \sim Ga\left(a + H - 1, b - \sum_{h=1}^{H-1} \log(1 - V_h)\right).$$

Multiple groups

Suppose we have Y_{ij} for $i = 1, \ldots, n_j$ and $j = 1, \ldots, J$, i.e. we have J groups with n_j observations per group. We may consider a DP for each group individually, i.e.

$$Y_{ij} \overset{ind}{\sim} \mathcal{G}_j, \quad \mathcal{G}_j \overset{ind}{\sim} DP(\alpha_j G_{0j})$$

where we must now specify α_j and G_{0j} for $j = 1, \ldots, J$. More importantly, this model does not allow us to borrow any information across the groups since the observations across groups given α_j and G_{0j}.

Some possible models to allow borrowing of information are the

- Dependent Dirichlet process (DDP)
- Hierarchical Dirichlet process (HDP)
- Nested Dirichlet process (NDP)

Dependent Dirichlet process

Suppose we are interested in estimating a collection of random probability measures $\mathcal{G}_1, \ldots, \mathcal{G}_J$. We would like for the measures to be DPs marginally, i.e.

$$\mathcal{G}_j \overset{ind}{\sim} DP(\alpha_j G_{0j})$$

but we may want to incorporate dependency between the measures and thus borrow information across the measures. One approach is a "fixed-π DDP" which is defined via the stick-breaking process such that each measure has the same weights π but the locations vary, i.e.

$$\mathcal{G}_j \overset{d}{=} \sum_{h=1}^{\infty} \pi_h \delta_{\theta^*_{jh}}, \quad \pi \sim \text{stick}(\alpha), \quad \theta^*_{jh} \sim G_0$$

Hierarchical Dirichlet process

An alternative is to build a hierarchical model, i.e.

$$\mathcal{G}_j \overset{ind}{\sim} DP(\alpha \mathcal{G}_0) \qquad \mathcal{G}_0 \sim DP(\beta G_{00})$$

The stick-breaking process related to this model is

$$\mathcal{G}_j \overset{d}{=} \sum_{h=1}^{\infty} \pi_{jh} \delta_{\theta_h^*}, \quad \mathcal{G}_0 \overset{d}{=} \sum_{h=1}^{\infty} \lambda_h \delta_{\theta_h^*}, \quad \theta_h^* \sim G_{00}$$

where

$$\pi_j = (\pi_{j1}, \pi_{j2}, \ldots) \sim \text{stick}(\alpha) \quad \text{and} \quad \lambda = (\lambda_1, \lambda_2, \ldots) \sim \text{stick}(\beta).$$

Like the DDP, the HDP allows individuals in different groups to be clustered together, i.e. have the same θ_h^*.

Nested Dirichlet process

Rather than clusting individuals across groups, we may be interested in clustering groups themselves, i.e. groups that have the same distribution should be treated as the same group. Here we can use the nested Dirichlet process:

$$\mathcal{G}_j \overset{ind}{\sim} \mathcal{G}, \quad \mathcal{G} \sim DP(\alpha \mathcal{G}_0), \quad \mathcal{G}_0 \equiv DP(\beta G_{00}).$$

The stick-breaking process related to this model is

$$\mathcal{G}_j \overset{ind}{\sim} \mathcal{G} \overset{d}{=} \sum_{h=1}^{\infty} \pi_h \delta_{\mathcal{G}_h^*}, \quad \pi \sim \text{stick}(\alpha), \quad \mathcal{G}_h^* \overset{ind}{\sim} DP(\beta G_{00}).$$

A natural combination of the HDP and NDP is to place a DP on G_{00} which results in common set of global atoms, i.e. θ_h^*, but with varying weights for each cluster.

Applications of DP

Primarily we have been discussing the use of the DP prior as a tool for Bayesian nonparametric density estimation. Here we discuss additional uses in the context of

- Random effects
- Error distributions
- Functional data analysis

Random effects model

Let y_{ij} be the observation for individual i in group j and assume

$$y_{ij} = \mu_j + \epsilon_{ij}, \quad \mu_j \overset{ind}{\sim} F, \quad \epsilon_{ij} \overset{ind}{\sim} G$$

A typical parametric model would assume $F \overset{d}{=} N(\eta, \tau^2)$ and $G \overset{d}{=} N(0, \sigma^2)$. Suppose we would like to be less informative about these distributional assumptions. One possibility is to assume

$$F \overset{ind}{\sim} DP(\alpha F_0).$$

Now we will estimate the density for the random effects μ_j. To estimate F, we will need many groups, i.e. J should be large. Alternatively (or additionally), we could assume

$$\epsilon_{ij} \overset{ind}{\sim} N(0, \sigma_i^2), \quad \sigma_i^2 \overset{ind}{\sim} G, \quad G \sim DP(\beta G_0).$$

Here we use the Dirichlet Process mixture to assure that the distribution for the observations are continuous. To estimate G, we need many observations per group.

Functional data analysis

Let y_{ij} be the observation and x_{ij} be an explanatory variable for individual i in group j and assume

$$y_{ij} = f_j(x_{ij}) + \epsilon_{ij}, \quad f_j(x) = \sum_{h=1}^{H} \theta_{jh} b_h(x)$$

where $b_h(x)$ for $h = 1, \ldots, H$ be a collection of basis functions. Now assume

$$\theta_j = (\theta_{j1}, \ldots, \theta_{jH}) \overset{ind}{\sim} G, \quad G \sim DP(\alpha G_0)$$

To provide parismony, i.e. dropping basis functions, we can utilize a base measure that a point-mass mixtures, i.e.

$$G_{0h} \overset{d}{=} \pi_{0h}\delta_0 + (1 - \pi_{0h})N(0, \tau_h^2)$$

If we want t alternatives, let $\tau_h^* \sim IG(\cdot, \cdot)$. A conditionally conjugate prior on the π is $\pi_{0h} \overset{ind}{\sim} Be(a, b)$. If exact zeros are not necessary, then let

$$\theta_{ch}^* \overset{ind}{\sim} N(0, \tau_{ch}^2), \quad \tau_{ch}^2 \overset{ind}{\sim} IG(\cdot, \cdot)$$

and thus have t distribution for the θ_{ch}^*, but now the MCMC is more efficient.

Bayesian nonparametrics in R

From CRAN Task View: Bayesian Inference, the packages that contain Dirichlet process related Bayesian nonparametrics are

- bayesm
- DPpackage
- growcurves
- PReMiuM

Density estimation in the DPpackage

```
library("DPpackage")
prior = list(alpha=1,
             m1 = 2,
             k0 = 1/3,
             nu1 = 0.2,
             psiinv1=diag(0.2,1))

mcmc = list(nburn=1000, nsave=10000, nskip=10, ndisplay=100)

state = NULL # initial state

dp = DPdensity(y = hat$phi,
               prior = prior,
               mcmc = mcmc,
               state=state,
               status=TRUE)
```

Density estimation in the DPpackage

```
?DPlmm
?DPglmm
?DPMlmm
?PMglmm
?DPolmm
?HDPMdensity

?PTdensity
```

References

- Neal, R. M. (2000). Markov Chain sampling methods for Dirichlet process mixture models. Journal of Computational and Graphical Statistics, 9: 249-265.