1 The models

In what follows, the three models that we are now considering are presented. The term $F(\cdot \mid \theta)$ and $H(\cdot \mid \mu)$ will denote general distribution functions with parameters (can be vectors) θ and μ, respectively, otherwise the distribution will be explicit.

Model 1:

$$
\begin{aligned}
& y_i \mid \theta_i \sim N(\theta_i, \sigma^2), \\
& \theta_i = \theta^\star_{\xi^\star_i}, \\
& \xi^\star_i \mid \boldsymbol{w} \sim Categorical(\boldsymbol{w}), \\
& w_k = z_k \prod_{l<k} (1 - z_l), \quad k = 2, \ldots, K, w_1 = z_1 \\
& z_k \sim Beta(1, \alpha), \quad k = 1, \ldots, K-1, z_K = 1, \\
& \theta^\star_k \sim N(\mu, \tau^2), \quad k = 1, \ldots, K.
\end{aligned}
$$

Model 2:

$$
\begin{aligned}
& y_i \mid \theta_i \sim F(\cdot \mid \theta_i), \\
& \theta_i = \tilde{\theta}_{\tilde{\xi}_i}, \\
& (\tilde{\xi}_1, \ldots, \tilde{\xi}_n) \sim CRP(\alpha), \\
& \alpha \sim H_1, \\
& \tilde{\theta}_k \sim H(\cdot \mid \mu), \quad k = 1, \ldots, \tilde{K}.
\end{aligned}
$$

Model 3:

$$
\begin{aligned}
& y_i \mid \theta_i \sim N(\theta_i, \sigma^2), \\
& \theta_i \mid G \sim G, \\
& G \sim DP(\alpha, N(\mu, \tau^2)).
\end{aligned}
$$

2 About Model 2

I'll give more details on the definition of the model in the BUGS code and describe the sampler. I will change the notation just to be consistent with the notation in code.

2.1 The model and its BUGS Code

The model:

$$
\begin{aligned}
&\tilde{\theta}_i \sim H(\cdot \mid \mu), \quad i = 1, \ldots, n,\\
&(\xi_1, \ldots, \xi_n) \sim CRP(\alpha),\\
&\alpha \sim H_1,\\
&\theta_i = \tilde{\theta}_{\xi_i},\\
&y_i \mid \theta_i \sim F(\cdot \mid \theta_i).
\end{aligned}
$$

BUGS code:

```
Code=nimbleCode({
    for(i in 1:n){
    thetatilde[i] ~ H
    }
    xi[1:n] ~ dCRP(conc)
    conc ~ H_1

    for(i in 1:n){
      theta[i] <- thetatilde[xi[i]]
      y[i] ~ F
    }
  })
```

I have tried different combinations of F and H: normal with known variance and normal, normal with unknown variance and normal, Poisson and gamma, Weibull and gamma.

Saying that $(\xi_1, \ldots, \xi_n) \sim CRP(\alpha)$ means that $\xi_1 = 1$, and

$$\xi_i \mid \xi_1, \ldots, \xi_{i-1} \sim \frac{1}{i-1+\alpha} \sum_{j=1}^{i-1} \delta_{\xi_j} + \frac{\alpha}{i-1+\alpha} \delta_{\xi^{new}},$$

where $\xi^{new} = \max\{\xi_1, \ldots, \xi_{i-1}\} + 1$.

2.2 Samplers

1. Sampling $\tilde{\theta}_k$: using NIMBLE's sampler.

2. Sampling ξ_i:

   (a) when H is not conjugate for F, a non conjugate sampler sampler based on algorithm 8 of Neal (2000) is used. More specifically, the ξ_i are updated one at the time from the following conditional distribution

$$\xi_i \mid \boldsymbol{y}, \xi_{-i}, ... \sim \frac{1}{n-1+\alpha} \sum_{j \neq i} f(y_i \mid \tilde{\theta}_{\xi_j}) \delta_{\xi_j} + \frac{\alpha}{n-1+\alpha} f(y_i \mid \tilde{\theta}_{\xi^{new}}) \delta_{\xi^{new}},$$

this is, ξ_i is an already existing label, ξ_j, or a new one, ξ^{new}, with probabilities proportional to $f(y_i \mid \tilde{\theta}_{\xi_j})$ and $\alpha f(y_i \mid \tilde{\theta}_{\xi^{new}})$, respectively, ξ^{new} is the smallest label related with no observations.

(b) when H is conjugate for F, we can integrate out $\tilde{\theta}$ eliminating them from the algorithm. In this case, the sampler is based on algorithm 3 of Neal (2000). Let $1, 2, \ldots, K$, be the relabeled K unique values in $(\xi_1, \ldots, \xi_n)$, let and $m_{-i,k}$ be the number of times label k appears in $\xi_{-i} = (\xi_1, \ldots, \xi_{i-1}, \xi_{i+1}, \ldots, \xi_n)$. The ξ_i are updated one at the time from the following conditional distribution

$$\xi_i \mid \boldsymbol{y}, \xi_{-i}, \ldots \sim \sum_{k=1}^{K} \frac{m_{-i,k}}{n-1+\alpha} f(y_i \mid \boldsymbol{y}_{-i,k}) \delta_k + \frac{\alpha}{n-1+\alpha} f(y_i) \delta_{k+1},$$

where $\boldsymbol{y}_{-i,k}$ denotes all observations y_l such that $l \neq i$ and $\xi_l = k$, $f(y_i \mid \boldsymbol{y}_{-i,k})$ is the posterior predictive density function at y_i based on data $\boldsymbol{y}_{-i,k}$ and the prior H, i.e., $f(y_i \mid \boldsymbol{y}_{-i,k}) = \int f(y_i \mid \theta) \prod_{\{j: j \neq i, \xi_j = k\}} f(y_j \mid \theta) H(d\theta)$, and $f(y_i)$ is the prior predictive density function at y_i, i.e., $f(y_i) = \int f(y_i \mid \theta) H(d\theta)$.

For instance, considering

- $F = N(\theta, \sigma^2)$, with known variance σ^2, and $H = N(\mu, \tau^2)$, it follows that $f(y_i \mid \boldsymbol{y}_{-i,k}) = N\left(y_i \mid \mu_{-i,1}, \sigma^2 + \tau^2_{-i,1}\right)$, where $\tau^2_{-i,1} = 1/\left(m_{-i,k}/\sigma^2 + 1/\tau^2\right)$, $\mu_{-i,1} = \tau^2_{-i,1} \left(\sum_{\{j: j \neq i, \xi_j = k\}} y_j/\sigma^2 + \mu/\tau^2\right)$, and $f(y_i) = N(y_i \mid \mu, \sigma^2 + \tau^2)$
- $F = Poisson(\theta)$ and $\theta \sim Gamma(a, b)$, it follows that $f(y_i \mid \boldsymbol{y}_{-i,k}) = \frac{b_1^{a_1}}{(b_1+1)^{a_1+y_i}} \frac{\Gamma(a_1+y_i)}{\Gamma(a_1)} \frac{1}{y_i!}$, where $a_1 = a + \sum_{\{j: j \neq i, \xi_j = k\}} y_j$, $b_1 = b + m_{-i,k}$, and $f(y_i) = \frac{b^a}{(b+1)^{a+y_i}} \frac{\Gamma(a+y_i)}{\Gamma(a)} \frac{1}{y_i!}$.

<u>Some inefficiencies</u>: 1) the *calculate* function is used at each step, 2) maybe we could update only few more $\tilde{\theta}$ than the unique ones, rather than the whole vector, in the random walk Metropolis step. (3) The ξ are updated one at the time, we could update only the ones related with observations (need of varying dimensions)).

A sampler for ξ in a conjugate case would update each ξ_i from the following conditional distribution

$$\xi_i \mid \boldsymbol{y}, \xi_{-i}, \ldots \sim \frac{1}{n-1+\alpha} \sum_{j \neq i} f(y_i \mid \tilde{\theta}_{\xi_j}) \delta_{\xi_j} + \frac{\alpha}{n-1+\alpha} f(y_i) \delta_{\xi^{new}},$$

where $f(y_i)$ is the prior predictive density function at y_i.

Is necessary to recognize what the predictive distribution from H and F is (bunch of if statements?).

3. Sampling α: if $H_1 \equiv Gamma(a, b)$, then the following sampler can be used:

   - sample $z \sim Beta(1 + \alpha, n)$ and compute $w = \frac{a+K-1}{a+K-1+n(b-ln(z))}$, where K is the number of unique values in $(\xi_1, \ldots, \xi_n)$.
   - sample $\alpha \sim Gama(a + K, b - ln(z))$ with probability w, and sample $\alpha \sim Gamma(a + K - 1, b - ln(z))$ with probability $1 - w$.

2.3 Output

Given samples $(\tilde{\theta}, (\xi_1, \ldots, \xi_n)$ or θ, we compute an approximation of measure G based on a truncation level, say L, given by the user.
Measure G is given by

$$G(\cdot) = \sum_{j=1}^{L} w_j \delta_{\theta_j^\star}(\cdot), \quad w_1 = v_1, \;\; w_j = v_j \prod_{l<j}(1 - v_l), l = 2, \ldots, L-1, \;\; w_L = \prod_{l<L}(1 - v_l),$$

where $v_l \sim Beta(1, \alpha + n)$, and

$$\theta_j^\star \sim \frac{\alpha}{\alpha + n} G_0 + \sum_{j=1}^{K} \frac{m_j}{\alpha + n} \delta_{\overline{\theta}_j},$$

where K is the number of unique values in $(\xi_1, \ldots, \xi_n)$, m_j denotes the number of relabeled $(\xi_1, \ldots, \xi_n)$ equal to j, and $\overline{\theta}_j$ denote the unique values in $(\tilde{\theta}_{\xi_1}, \ldots, \tilde{\theta}_{\xi_n})$.

Comments: K, m_j, and $\overline{\theta}$ can be obtained from θ.
There are results relating α, and acceptable error, ϵ, and the truncation level, L, of G. More specifically, $\left(\frac{\alpha}{\alpha+1}\right)^{L-1} = \epsilon$, so we can give a warning when the truncation level is too small for the value (or samples) of α and ϵ.

3 Model 1

We don't have a BUGS version for this model as is stated before. We have BUGS code for another representation of this model that involves the random measure G written as the truncation of its stick breaking representation written as a matrix, and integrating out the random indexes.

3.1 The model and its BUGS Code

The model:

$$
\begin{aligned}
& w_k = z_k \prod_{l<k}(1-z_l), \quad k = 2, \dots, T, w_1 = z_1 \\
& z_k \sim Beta(1, \alpha), \quad k = 1, \dots, T-1, z_T = 1, \\
& \theta^\star_k \sim N(\mu, \tau^2), \quad k = 1, \dots, T, \\
& \theta_i = \theta^\star_{\xi^\star_i}, \\
& \xi^\star_i \mid \boldsymbol{w} \sim Categorical(\boldsymbol{w}), \\
& y_i \mid \theta_i \sim N(\theta_i, \sigma^2).
\end{aligned}
$$

BUGS code:

```
Code=nimbleCode(
  {
    G[1:T,1:2] ~ dtruncSBDPnormal(
        conc=conc0, mean=mean0, sd=tau0)
    theta[1:T] ~ dNPDiscreteV(G[1:T,1:2])
    for(i in 1:n){
      y[i] ~ dnorm(theta[i], sd=s0)
    }
    conc0 <-1;
    mean0 <- 5; tau0 <- sqrt(10);
     s0 <- sqrt(10)
})
```

For this model only the conjugate normal-normal model has been considered.

In the BUGS code, saying `G[1:T,1:2] ~ truncSBDPnormal(conc=conc0, mean=mean0, sd=tau0)` means that

$$G(\cdot) = \sum_{k=1}^{T} w_k \delta_{\theta_k}(\cdot), \quad w_k = z_k \prod_{l<k}(1-z_l), \;\; z_k \sim Beta(1, \alpha), \;\; \theta_k \sim N(\mu, \tau^2),$$

and saying $theta[1:T] \; dNPDiscreteV(G[1:T, 1:2])$ means that

$$\theta_k \sim \sum_{j=1}^{T} w_j \delta_{m_j},$$

where w_j are weights that add up to one, and m_j are possible values for θ_k.

Note that lines 1 to 3 in the model correspond to `G[1:T,1:2] ~ dtruncSBDPnormal(conc=conc0, mean=mean0, sd=tau0)` in the BUGS code, and lines 4 and 5 are related to `theta[1:T] ~ dNPDiscreteV(G[1:T,1:2])` in the BUGS code.

3.2 Sampler

The sampler that we have is the blocked Gibbs sampler and samples the random measure G as a matrix whose first and second columns are the vector of $\theta^\star$ and w, respectively. The steps of the sampler are the following:

a) updating $(\theta_1^\star, \ldots, \theta_T^\star)$: $\theta_k^\star$ is updated from the prior if it is no related with any observation. Otherwise, $\theta_k^\star$ is updated from the posterior (which is conjugate in this case) considering the observations that have label k.

b) updating the labels $\xi^\star$:

$$\xi_i^\star \mid \boldsymbol{y}, ... \sim \sum_{k=1}^{T} w_k N(y_i \mid \theta_k^\star, \sigma^2)\delta_k.$$

c) updating the weights $(w_1, \ldots, w_T)$: first the stick variables are updated as

$$z_k \mid \boldsymbol{y}, ... \sim Beta\left(1 + M_k, \alpha + \sum_{l=k+1}^{T} M_l\right),$$

where M_k denotes the number of observations that have label k. Then $w_k = v_k \prod_{l<k}(1 - z_l)$.

Several changes: 1) $\theta^\star$ could be updated using NIMBLE samplers, 2) I need to check the conjugacy between z_k, w_k and $\xi^\star$ (Dirichlet-categorical), 3) if there is conjugacy, then use the NIMBLE sampler too, and 4) create a sampler only for the $\xi^\star$.

References

NEAL, R. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* 9 249–265.