Computational methods for Bayesian nonparametric models

Jim Griffin
University College London

Structure

- Pólya urn scheme samplers – Conjugate/non-conjugate DPMM, split-and-merge.
- Stick-breaking – Block Gibbs sampler, slice sampler
- Normalized random measures – Simulation, Truncation methods, slice sampling

Not covered

- Feature allocation – Indian buffet process, beta-Bernoulli processes and generalisations
- More complicated processes – Poisson-Kingman processes, etc.
- Distributed computing
- Sequential Monte Carlo
- Variational Bayes
- More complicated models – hierarchical processes, nested processes, dependent processes, etc.

Bayesian nonparametric mixture models

Hierarchical model

$$y_i|\theta_i \sim k(y_i \mid \theta_i), \quad \theta_i|F \sim F, \quad F \sim \mathrm{DP}(M, H)$$

Allocation representation

$$y_i|s_i, \theta_1^\star, \theta_2^\star, \cdots \sim k\left(y_i|\theta_{s_i}^\star\right),$$

$$p(s_i = k) = w_k \sim \mathrm{GEM}(M)^{1}, \quad \theta_k^\star \overset{i.i.d.}{\sim} H$$

where $\theta_1^\star, \theta_2^\star, \ldots$ are the distinct values

[1] Griffiths-Engen-McCloskey process

Marginal MCMC methods for Dirichlet process mixture models

The key idea here is that the predictive distribution of the allocation variable s_j is described by a Chinese restaurant process

$$p(s_j = k | s_1, \ldots, s_{n-1}) = \begin{cases} \frac{n_{k,j}}{M+j-1} & 1 \leq k \leq K_{j-1} \\ \frac{M}{M+j-1} & k = K_{j-1} + 1 \end{cases}$$

where K_{j-1} is the number of different values of $s_1, \ldots, s_{j-1}$ and $n_{k,j}$ is the number of observations in the k-th cluster.

Then

$$p(s_1, \ldots, s_n) = \prod_{j=1}^{n} p(s_j | s_1, \ldots, s_{j-1})$$

Marginal MCMC methods for Dirichlet process mixture models

The posterior is finite (but not fixed) dimensional

$$p(s_1,\ldots,s_n,\theta_1^\star,\ldots,\theta_{K_n}^\star|y)\propto\prod_{k=1}^{K_n}\prod_{\{i|s_i=k\}}k(y_i|\theta_k^\star)\prod_{k=1}^{K_n}h(\theta_k^\star)p(s_1,\ldots,s_n)$$

Marginal MCMC methods for Dirichlet process mixture models

The posterior is finite (but not fixed) dimensional

$$p(s_1,\ldots,s_n,\theta_1^\star,\ldots,\theta_{K_n}^\star|y)\propto\prod_{k=1}^{K_n}\prod_{\{i|s_i=k\}}k(y_i|\theta_k^\star)\prod_{k=1}^{K_n}h(\theta_k^\star)p(s_1,\ldots,s_n)$$

In a conjugate DPM we assume that $\int\prod_{\{i|s_i=k\}}k(y_i|\theta)h(\theta)\,d\theta$ can be calculate analytically.

$$p(s_1,\ldots,s_n|y)=\int p(s_1,\ldots,s_n,\theta_1^\star,\ldots,\theta_{K_n}^\star|y)\,d\theta_1^\star\ldots d\theta_{K_n}^\star$$

Marginal MCMC methods for Dirichlet process mixture models

The posterior is finite (but not fixed) dimensional

$$p(s_1,\dots,s_n,\theta^\star_1,\dots,\theta^\star_{K_n}|y) \propto \prod_{k=1}^{K_n}\prod_{\{i|s_i=k\}} k(y_i|\theta^\star_k)\prod_{k=1}^{K_n} h(\theta^\star_k)p(s_1,\dots,s_n)$$

In a conjugate DPM we assume that $\int\prod_{\{i|s_i=k\}} k(y_i|\theta)h(\theta)\,d\theta$ can be calculate analytically.

Then

$$p(s_1,\dots,s_n|y) = \int p(s_1,\dots,s_n,\theta^\star_1,\dots,\theta^\star_{K_n}|y)\,d\theta^\star_1\dots d\theta^\star_{K_n}$$

$$\propto \prod_{k=1}^{K_n}\int\prod_{\{i|s_i=k\}} k\left(y_i|\theta^\star\right)\,h\left(\theta^\star_k\right)\,d\theta^\star_k \times p(s_1,\dots,s_n)$$

Gibbs sampler

Due to exchangeability

$$p(s_j = k|s_1, \ldots, s_{n-1}) = \begin{cases} \frac{n_k^{-j}}{M+n-1} k(y_j|\{y_i|s_i = k\}) & 1 \leq k \leq K^{-j} \\ \frac{M}{M+n-1} k(y_j) & k = K^{-j} + 1 \end{cases}$$

where

- $n_k^{-j} = \#\{s_i = k, i \neq j\}$.
- $k(y_j|\{y_j|s_i = k\})$ is the predictive distribution with likelihood k and prior H
- K^{-j} is the number of distinct values in $s_1, \ldots, s_{j-1}, s_{j+1}, \ldots, s_n$.

Non-conjugate Dirichlet process mixtures – Neal's algorithm 8

We work on $s_1, \ldots, s_n$ and $\theta_1^\star, \ldots, \theta_{K_n}^\star$. The posterior is proportional to

$$p(s_1, \ldots, s_n) \prod_{i=1}^{n} k(y_i | \theta_{s_i}^\star) \prod_{k=1}^{K_n} h(\theta_k^\star)$$

Non-conjugate Dirichlet process mixtures – Neal's algorithm 8

We work on $s_1, \ldots, s_n$ and $\theta_1^\star, \ldots, \theta_{K_n}^\star$. The posterior is proportional to

$$p(s_1, \ldots, s_n) \prod_{i=1}^{n} k(y_i|\theta_{s_i}^\star) \prod_{k=1}^{K_n} h(\theta_k^\star)$$

where

$$p(s_1, \ldots, s_n) = M^{K_n} \frac{\Gamma(n)}{\Gamma(M+n)} \prod_{k=1}^{K_n} \Gamma(n_k)$$

and $n_k = \#\{s_i = k\}$. This is the exchangeable product partition formula (EPPF).

Neal's algorithm 8

In this sampler, latent variables $\psi_1, \ldots, \psi_m$ are introduced to define the augmented posterior (by exchangeability, we only look at s_n).

Neal's algorithm 8

In this sampler, latent variables $\psi_1, \ldots, \psi_m$ are introduced to define the augmented posterior (by exchangeability, we only look at s_n).

$$M^{K^{-n}} \prod_{k=1}^{m} \left(\frac{M}{m}\right)^{\mathrm{I}(s_n = K_{-n}+1 \text{ and } \theta^\star_{K_{-n}+1} = \psi_k)}$$
$$\times \frac{\Gamma(n)}{\Gamma(M+n)} \prod_{k=1}^{K_n} \Gamma(n_k) \prod_{i=1}^{n} k(y_i | \theta^\star_{s_i}) \prod_{k=1}^{K_{-n}} h(\theta^\star_k) \prod_{k=1}^{m} h(\psi_k)$$

Full conditional distribution

To update the allocation variable for the i-th observations re-label the s's so that $s_i \geq s_j$ for $1 \leq j \leq n$.

Full conditional distribution

To update the allocation variable for the *i*-th observations re-label the *s*'s so that $s_i \geq s_j$ for $1 \leq j \leq n$.

The full conditional distribution of s_i is

$$p(s_i = k) \propto \begin{cases} n_{i,k}\, k(y_i|\theta^\star_k) & 1 \leq k \leq K_{-n} \\ \frac{M}{m} k(y_i|\psi_{k-K_{-n}}) & K_{-n}+1 \leq k \leq K_{-n}+m \end{cases}$$

where

- If the *i*-th observation is a singleton (*i.e.* $n_{s_i} = 1$) then $\psi_1 = \theta^\star_{K_{-n}+1}$ and $\psi_2, \ldots, \psi_m \sim H$.
- Otherwise, $\psi_1, \ldots, \psi_m \sim H$.

Full conditional distribution

To update the allocation variable for the i-th observations re-label the s's so that $s_i \geq s_j$ for $1 \leq j \leq n$.

The full conditional distribution of s_i is

$$p(s_i = k) \propto \begin{cases} n_{i,k}\, k(y_i|\theta^\star_k) & 1 \leq k \leq K_{-n} \\ \frac{M}{m} k(y_i|\psi_{k-K_{-n}}) & K_{-n}+1 \leq k \leq K_{-n}+m \end{cases}$$

where

- If the i-th observation is a singleton ($i.e.$ $n_{s_i} = 1$) then $\psi_1 = \theta^\star_{K_{-n}+1}$ and $\psi_2, \ldots, \psi_m \sim H$.
- Otherwise, $\psi_1, \ldots, \psi_m \sim H$.

If $k > K_{-n}$, set $s_n = K_{-n}+1$ and $\theta^\star_{K_{-n}+1} = \psi_{k-K_{-n}}$.

Comments

- Algorithm 8 has been extended to the case of normalized random measures with independent increments by Favaro and Teh (2013).
- The approach can be interpreted as akin to a pseudo-marginal method (Andrieu and Roberts, 2009).

Split-merge samplers

Split-merge samplers try to make larger moves in allocation space.

Two approaches SAMS (Dahl and Newcomb, 2022) and RGMS (Neal and Jian, 2004) which differ in re-allocation mechanism.

Both algorithms involve sampling two observations i and j at random.

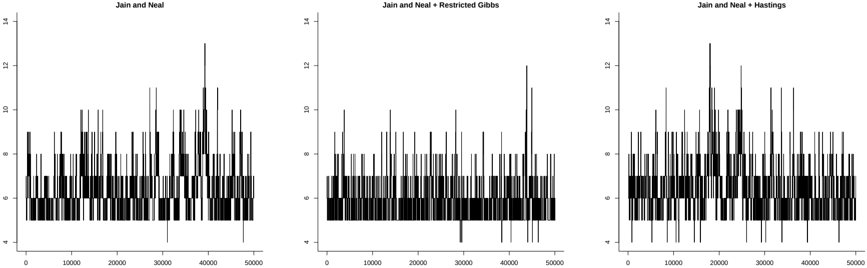
- If these $s_i \neq s_j$, merge.
- If there $s_i = s_j$, split. Create two clusters containing observation i and j respectively.
  - SAMS – Allocate all other observations with $s_k = s_i$ i one of the two new clusters using the full conditional (sequentially).
  - RGMS – Randomly allocate all observations with $s_k = s_i$ to the two clusters. Update the allocation using the full conditional distribution.

Joint work with Alessandro Colombi

These can be seen as a type of informed neighbourhood scheme where

- The neighbourhood is all possible re-clusterings of the two clusters conditional on the anchor points.
- The sequential/restricted Gibbs scan is a type of informed proposal.

We can build an adaptive random neighbourhood informed (ARNI) scheme where the random neighbourhood is a core-periphery model with a ‘‘score" of ‘‘core-ness” for each observation.


Jain and Neal
Jain and Neal + Restricted Gibbs
Jain and Neal + Hastings
4
6
8
10
12
14
0
10000
20000
30000
40000
50000

Conditional samplers for DPM models

Conditional samplers use a particular representation of a Dirichlet process and choose a truncation level.

Conditional samplers for DPM models

Conditional samplers use a particular representation of a Dirichlet process and choose a truncation level.

Usually, these algorithms work for both conjugate and non-conjugate DPM models.

Representations of the Dirichlet process

Suppose that $F \sim \mathrm{DP}(M, H)$ then

$$F = \sum_{k=1}^{\infty} w_k \delta_{\theta_k}$$

where $\theta_k \sim H$ and ...

Representations of the Dirichlet process

Suppose that $F \sim \mathrm{DP}(M, H)$ then

$$F = \sum_{k=1}^{\infty} w_k \delta_{\theta_k}$$

where $\theta_k \sim H$ and ...

1. Stick-breaking representation
   $w_k = V_k \prod_{j<k}(1 - V_j)$ where $V_j \overset{i.i.d.}{\sim} \mathrm{Be}(1, M)$.

Representations of the Dirichlet process

Suppose that $F \sim \mathrm{DP}(M, H)$ then

$$F = \sum_{k=1}^{\infty} w_k \delta_{\theta_k}$$

where $\theta_k \sim H$ and ...

1. Stick-breaking representation
   $w_k = V_k \prod_{j<k}(1 - V_j)$ where $V_j \overset{i.i.d.}{\sim} \mathrm{Be}(1, M)$.
2. Normalized random measure with independent increment
   $w_k = \frac{\gamma_k}{\sum_{m=1}^{\infty} \gamma_m}$ where $\gamma_1, \gamma_2, \ldots$ are the jumps of a gamma process.

Samplers using stick-breaking representations

Choose a value K and set $V_{K+1} = 1$ then only $w_1, \ldots, w_{K+1}$ are non-zero.

Samplers using stick-breaking representations

Choose a value K and set $V_{K+1} = 1$ then only $w_1, \ldots, w_{K+1}$ are non-zero.

The truncation error can be measured in the following way

$$\| f(y) - f_K(y) \|_1 < 4 \left(1 - \mathrm{E}\left[\left(\sum_{k=1}^{K} w_k \right)^n \right] \right)$$

where

- $f(y)$ is the joint predictive distribution of the data
- $f_K(y)$ is the joint predictive distribution of the data under the truncation.
- $\| \cdot \|_1$ represents L_1 distance.

Blocked Gibbs sampler (Ishwaran and James, 2001)

If we truncate at K components, then

$$p(s_1, s_2, \ldots, s_n) = \prod_{i=1}^{n} w_{s_i} = \prod_{k=1}^{K} V_k^{n_k}(1 - V_k)^{m_k}$$

where $n_k = \#\{i|s_i = k\}$ and $m_k = \sum_{i=k+1}^{K} n_i$.

Blocked Gibbs sampler

1. The full conditional for s_i is

$$p(s_i = k) \propto w_k \, k(y_i|\theta_k), \qquad k = 1, \ldots, K+1$$

2. The full conditional for V_k is $\text{Be}(1 + n_k, M + m_k)$.
3. The full conditional for θ_k is proportional to

$$h(\theta_k) \prod_{\{i|s_i=k\}} k(y_i|\theta_k)$$

Note: This full conditional distribution will have a known if the prior distribution of θ_k is conjugate.

Comments

- This method can be directly applied to the Pitman-Yor process where $V_k \sim \text{Be}(1-a, M+k\,a)$.
- The form $p(s_1, s_2, \ldots, s_n) = \prod_{k=1}^{K} V_k^{n_k}(1-V_k)^{m_k}$ is convenient for other models such as probit stick-breaking (Rodriguez and Dunson, 2011)

A slice sampler for DPM models

The truncation method leads to a simple Gibbs sampler but involves a truncation error.

A slice sampler for DPM models

The truncation method leads to a simple Gibbs sampler but involves a truncation error.

The slice sampler uses a random truncation to avoid a truncation error.

Slice sampler for DPM models

In the infinite-dimensional model, the marginal distribution of y_i is

$$p(y_i) = \sum_{k=1}^{\infty} w_k \, k(y_i|\theta_k)$$

Slice sampler for DPM models

In the infinite-dimensional model, the marginal distribution of y_i is

$$p(y_i) = \sum_{k=1}^{\infty} w_k \, k(y_i|\theta_k)$$

The slice sampler introduces a latent variable u_i

$$p(y_i, u_i) = \sum_{k=1}^{\infty} \mathrm{I}(u_i < w_k) \, k(y_i|\theta_k)$$

Slice sampler for DPM models

In the infinite-dimensional model, the marginal distribution of y_i is

$$p(y_i) = \sum_{k=1}^{\infty} w_k \, k(y_i|\theta_k)$$

The slice sampler introduces a latent variable u_i

$$p(y_i, u_i) = \sum_{k=1}^{\infty} \mathrm{I}(u_i < w_k) \, k(y_i|\theta_k)$$

Note: The marginal of y_i is unchanged.

Slice sampler for DPM models

$$p(y_i, u_i, s_i) = \mathrm{I}(u_i < w_{s_i})\, k(y_i|\theta_{s_i})$$

Note: The marginal of y_i is unchanged.

Slice sampler for DPM models

$$p(y_i, u_i, s_i) = \mathrm{I}(u_i < w_{s_i})\, k(y_i|\theta_{s_i})$$

Note: The marginal of y_i is unchanged.

Slice sampler for DPM models

$$p(y_i, u_i, s_i) = \mathrm{I}(u_i < w_{s_i})\, k(y_i|\theta_{s_i})$$

Note: The marginal of y_i is unchanged.

The posterior distribution is

$$p(u, s, V, \theta|y) \propto h(\theta)\, p(V) \prod_{i=1}^{n} \mathrm{I}(u_i < w_{s_i})\, k(y_i|\theta_{s_i})$$

Gibbs sampler (Walker, 2007)

1. Full conditional distribution of s_i

$$p(s_i = k) \propto \mathrm{I}(w_k > u_i)\, k(y_i|\theta_k)$$

   Note: there are only a finite number of non-zero values.

2. Full conditional of θ_k is

$$h(\theta_k) \prod_{\{i|s_i=k\}} k(y_i|\theta_k)$$

3. Full conditional distribution of $u_i \sim \mathrm{U}(0, w_{s_i})$.
4. Full conditional distribution of V_k is proportional to

$$p(V_k) \prod_{i=1}^{n} I\left(u_i < V_{s_l} \prod_{j<s_i} (1 - V_j)\right).$$

Accelerated Gibbs sampler (Kalli et al., 2011)

It is more efficient to jointly updating $v_1, \ldots, v_K$ and $u_1, \ldots, u_n$ using

$$p(v_1, \ldots, v_K, u_1, \ldots, u_n) = p(u_1, \ldots, u_n | v_1, \ldots, v_K) p(v_1, \ldots, v_K)$$

Accelerated Gibbs sampler (Kalli et al., 2011)

It is more efficient to jointly updating $v_1, \ldots, v_K$ and $u_1, \ldots, u_n$ using

$$p(v_1, \ldots, v_K, u_1, \ldots, u_n) = p(u_1, \ldots, u_n | v_1, \ldots, v_K) p(v_1, \ldots, v_K)$$

Then steps 3 and 4 become

1. The full conditional for V_k is $\text{Be}(1 + n_k, M + m_k)$. *i.e.* full conditional from blocked Gibbs sampler.
2. Full conditional distribution of $u_i \sim \text{U}(0, w_{s_i})$.

To complete step 1, we only need $w_1, \ldots, w_K > \min\{u_i\}$.

To complete step 1, we only need $w_1, \ldots, w_K > \min\{u_i\}$.

Notice that $w_k < \prod_{j=1}^{K}(1 - V_j)$ for $k > K$ and that, for $k > \max\{s_i\}$, $V_k \sim \text{Be}(1, M)$.

To complete step 1, we only need $w_1, \ldots, w_K > \min\{u_i\}$.

Notice that $w_k < \prod_{j=1}^{K}(1 - V_j)$ for $k > K$ and that, for $k > \max\{s_i\}$, $V_k \sim \text{Be}(1, M)$.

So we choose the first K for which $\prod_{j=1}^{K}(1 - V_j) < \min\{u_i\}$ which implies that $w_k < \min\{u_i\}$ for $k > K$.

Completely random measures

$\tilde{\mu}$ is a completely random measure (CRM) on $\mathbb{X}$ if, for any disjoint subsets $A_1, \ldots, A_n$, $\tilde{\mu}(A_1) \ldots, \tilde{\mu}(A_n)$ are mutually independent.

Completely random measures

$\tilde{\mu}$ is a completely random measure (CRM) on $\mathbb{X}$ if, for any disjoint subsets $A_1, \ldots, A_n$, $\tilde{\mu}(A_1) \ldots, \tilde{\mu}(A_n)$ are mutually independent.

We concentrate on completely random measures (CRM's) which can be represented in terms of jump sizes J_i and jump locations X_i as

$$\tilde{\mu} = \sum_{i=1}^{\infty} J_i \delta_{X_i}$$

where δ is Dirac's delta function and have Lévy-Khintchine representation

$$\mathbb{E}\left[e^{-\int f(x)\tilde{\mu}(dx)}\right] = e^{-\int_0^{\infty}\int \left[1 - e^{-sf(x)}\right] h(dx)\nu(ds)}$$

where $h(\cdot)$ is a p.d.f. and $\nu(\cdot)$ is a Lévy intensity $\int_{\mathbb{R}\setminus\{0\}} \min(1, x^2)\nu(x)\,dx < \infty$.

Simulating completely random measures

Suppose that we wish to simulate a realisation of a Lévy process with intensity ν.

Simulating completely random measures

Suppose that we wish to simulate a realisation of a Lévy process with intensity ν.

Define the tail mass function η of a Lévy process with Lévy intensity ν to be $\eta(x) = \int_x^\infty \nu(z)\, dz$.

Simulating completely random measures

Suppose that we wish to simulate a realisation of a Lévy process with intensity ν.

Define the tail mass function η of a Lévy process with Lévy intensity ν to be $\eta(x) = \int_x^\infty \nu(z)\,dz$.

Ferguson and Klass (1972) showed the Lévy process can be represented as

$$F = \sum_{k=1}^{\infty} J_k \delta_{\theta_k}$$

where $\gamma_k = \eta^{-1}(E_j)$ and $E_1, E_2, \ldots$ are the points of a Poisson process with intensity 1.

Example: Gamma process

Recall that a gamma process has Lévy intensity $\nu(x) = Mx^{-1} \exp\{-x\}$.

Example: Gamma process

Recall that a gamma process has Lévy intensity $\nu(x) = Mx^{-1} \exp\{-x\}$.

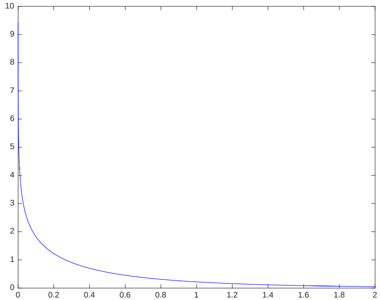
The tail mass integral is

$$\eta(x) = M \int_x^{\infty} z^{-1} \exp\{-z\}\, dz = M\, \mathrm{Ei}(x)$$

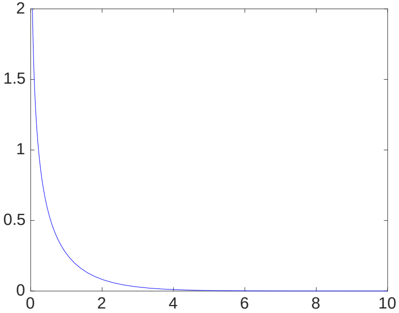
where $\mathrm{Ei}(x)$ is the exponential-integral function.

Gamma process simulation

$\eta(x)$

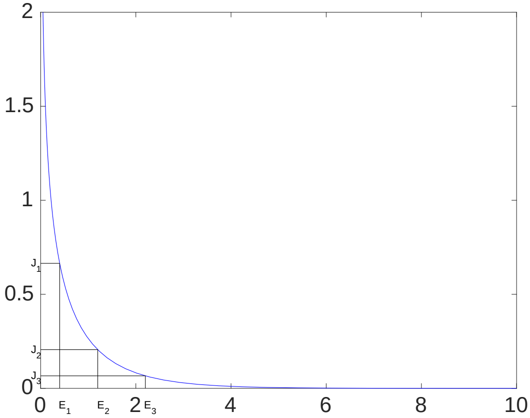


$\eta^{-1}(x)$

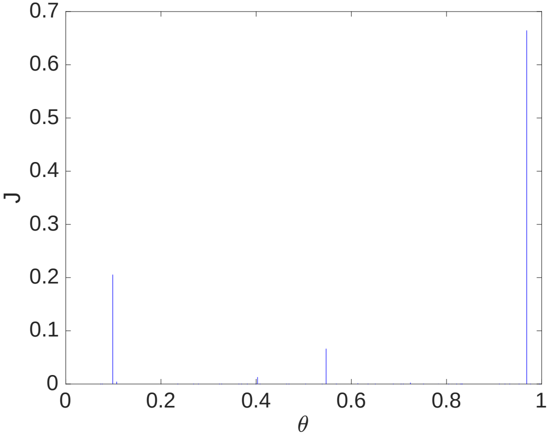


Gamma process simulation

$\eta^{-1}(x)$



Gamma process simulation



In general, the Ferguson-Klass algorithm can be slow since $\eta(x)$ can only be numerically inverted for the gamma process.

Rejection method (Rosinski, 2001)

Suppose that we have two non-Gaussian Lévy process with Lévy intensities $\nu_1(x)$ and $\nu_2(x)$ with $\nu_1(x) < \nu_2(x)$ for all x.

Rejection method (Rosinski, 2001)

Suppose that we have two non-Gaussian Lévy process with Lévy intensities $\nu_1(x)$ and $\nu_2(x)$ with $\nu_1(x) < \nu_2(x)$ for all x.

We can simulate a sequence $z_1, z_2, \ldots$ with intensity $\nu_1(x)$ by thinning a sequence $y_1, y_2, \ldots$ with intensity $\nu_2(x)$.

Rejection method (Rosinski, 2001)

Suppose that we have two non-Gaussian Lévy process with Lévy intensities $\nu_1(x)$ and $\nu_2(x)$ with $\nu_1(x) < \nu_2(x)$ for all x.

We can simulate a sequence $z_1, z_2, \ldots$ with intensity $\nu_1(x)$ by thinning a sequence $y_1, y_2, \ldots$ with intensity $\nu_2(x)$.

Set $i = 1$ and $j = 1$

1. Simulate y_i from $\nu_2(x)$ using the Ferguson-Klass algorithm.

Rejection method (Rosinski, 2001)

Suppose that we have two non-Gaussian Lévy process with Lévy intensities $\nu_1(x)$ and $\nu_2(x)$ with $\nu_1(x) < \nu_2(x)$ for all x.

We can simulate a sequence $z_1, z_2, \dots$ with intensity $\nu_1(x)$ by thinning a sequence $y_1, y_2, \dots$ with intensity $\nu_2(x)$.

Set $i = 1$ and $j = 1$

1. Simulate y_i from $\nu_2(x)$ using the Ferguson-Klass algorithm.
2. Simulate a uniform random variable u and calculate $\omega_i = \nu_1(y_i)/\nu_2(y_i)$. If $u < \omega_i$ set $z_j = y_i$, $j = j + 1$, $i = i + 1$. Otherwise, set $i = i + 1$.

Rejection method (Rosinski, 2001)

Suppose that we have two non-Gaussian Lévy process with Lévy intensities $\nu_1(x)$ and $\nu_2(x)$ with $\nu_1(x) < \nu_2(x)$ for all x.

We can simulate a sequence $z_1, z_2, \dots$ with intensity $\nu_1(x)$ by thinning a sequence $y_1, y_2, \dots$ with intensity $\nu_2(x)$.

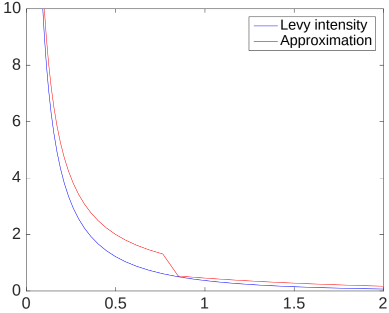
Set $i = 1$ and $j = 1$

1. Simulate y_i from $\nu_2(x)$ using the Ferguson-Klass algorithm.
2. Simulate a uniform random variable u and calculate $\omega_i = \nu_1(y_i)/\nu_2(y_i)$. If $u < \omega_i$ set $z_j = y_i$, $j = j + 1$, $i = i + 1$. Otherwise, set $i = i + 1$.
3. Goto step 1.

Bounding functions for the gamma process (Griffin, 2019)

$$\nu_2(x) = \begin{cases} -\frac{1}{x} & x < b \\ b^{-1}\exp\{-x\} & x \geq b \end{cases},$$

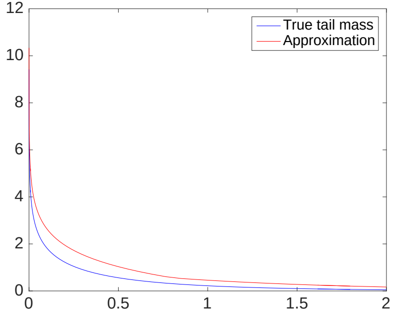
where $b = 0.8065$.



Bounding functions for the gamma process

$$\eta_2(x) = \begin{cases} -\log x + \log b + b^{-1} \exp\{-b\} & x < b \\ b^{-1} \exp\{-x\} & x \geq b \end{cases},$$

where $b = 0.8065$.



Truncation methods for gamma processes

1. Normalized gamma random variables
   $w_j = \frac{\gamma_j}{\sum_{k=1}^{K} \gamma_k}$ where $\gamma_k \sim \text{Ga}(M/K, 1)$.

Truncation methods for gamma processes

1. Normalized gamma random variables
   $w_j = \frac{\gamma_j}{\sum_{k=1}^{K} \gamma_k}$ where $\gamma_k \sim \mathrm{Ga}(M/K, 1)$.
2. Ferguson and Klass representation
   Choose a finite number of jumps K

$$F = \sum_{k=1}^{K} J_k \delta_{\theta_k}$$

Note: the jumps are ordered $J_1 > J_2 > J_3 > \dots$.

3. ϵ-approximation

$$F = J_0\delta_{\theta_0} + \sum_{k=1}^{K_\epsilon} J_k\delta_{\theta_k}$$

where $J_1, \dots, J_{K_\epsilon}$ are all jumps greater than ϵ.

$K_\epsilon \sim \text{Pn}(\int_\epsilon^\infty \nu(x)\,dx)$(also, $K_\epsilon \sim \text{Pn}(\eta(\epsilon))$) and

$$p(J_k) = \frac{\nu(J_k)}{\int_\epsilon^\infty \nu(x)\,dx}, \qquad J_k > \epsilon \text{ for } k = 0, 1, \dots, K_\epsilon.$$

Note: $\sum_{k=1}^{K_\epsilon} J_k\delta_{\theta_k}$ is a compound Poisson process and the truncation error (without J_0) is a Lévy process.

Illustration of ϵ-truncation

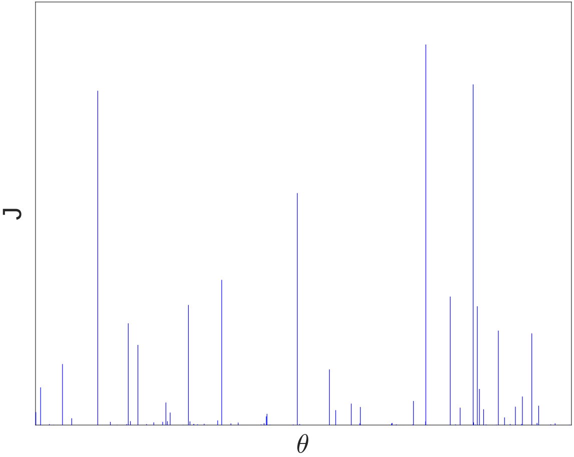


Illustration of ϵ-truncation

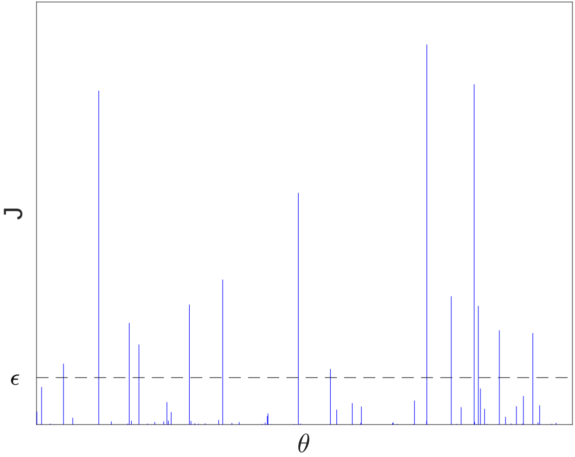
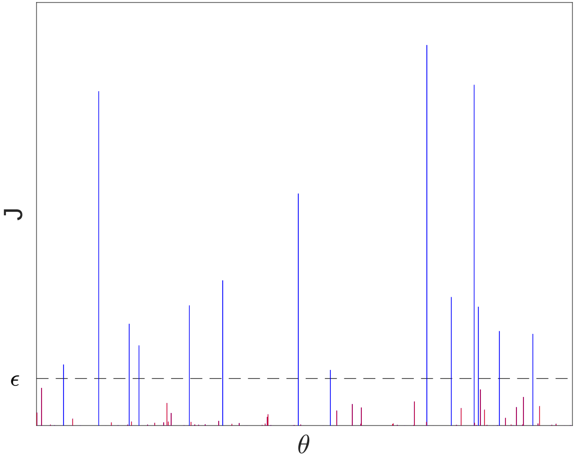


Illustration of ϵ-truncation



Sampling using normalized gamma random variables

The posterior distribution of $s_1, \ldots, s_n, \gamma_1, \ldots, \gamma_K, \theta_1, \ldots, \theta_K$ is proportional to

$$\prod_{k=1}^{K} p(\gamma_k) \prod_{k=1}^{K} h(\theta_k) \prod_{i=1}^{n} w_{s_i} \prod_{i=1}^{n} k(y_i|\theta_{s_i})$$

Sampling using normalized gamma random variables

The posterior distribution of $s_1, \ldots, s_n, \gamma_1, \ldots, \gamma_K, \theta_1, \ldots, \theta_K$ is proportional to

$$\prod_{k=1}^{K} p(\gamma_k) \prod_{k=1}^{K} h(\theta_k) \prod_{i=1}^{n} w_{s_i} \prod_{i=1}^{n} k(y_i|\theta_{s_i})$$

$$= \prod_{k=1}^{K} p(\gamma_k) \prod_{k=1}^{K} h(\theta_k) \frac{\prod_{i=1}^{n} \gamma_{s_i}}{\left(\sum_{j=1}^{K} \gamma_j\right)^n} \prod_{i=1}^{n} k(y_i|\theta_{s_i})$$

Sampling using normalized gamma random variables

The "e to the minus v" trick

$$\frac{1}{\Gamma(n)}\int v^{n-1}\exp\{-va\}\,dx=\frac{1}{a^n}$$

Sampling using normalized gamma random variables

The "e to the minus v" trick

$$\frac{1}{\Gamma(n)}\int v^{n-1}\exp\{-va\}\,dx = \frac{1}{a^n}$$

We define a posterior augmented with v which is proportional to

$$\prod_{k=1}^{K} p(\gamma_k)\prod_{k=1}^{K} h(\theta_k)\prod_{i=1}^{n}\gamma_{s_i} v^{n-1}\exp\left\{-v\sum_{j=1}^{K}\gamma_j\right\}\prod_{i=1}^{n} k(y_i|\theta_{s_i})$$

The Gibbs sampler

The full conditional of the Gibbs sampler are

1. The full conditional distribution of v is $\text{Ga}(n, \sum_{j=1}^{K} \gamma_j)$.
2. The full conditional distribution of γ_k is $\text{Ga}(M/K + n_k, 1 + v)$.
3. The full conditional distribution of s_i is a discrete distribution with

$$p(s_i = k) \propto \gamma_k \, k(y_i|\theta_k)$$

4. The full conditional distribution of θ_k has density proportional to

$$h(\theta_k) \prod_{\{i|s_i=k\}} k(y_i|\theta_{s_i})$$

Sampling using ϵ-truncation

An augment posterior is constructed in a similar way to before

$$p(v, s_1, \ldots, s_n, J_0, \ldots, J_{K_\epsilon}, \theta_0, \ldots, \theta_{K_\epsilon})$$
$$\propto \prod_{k=0}^{K_\epsilon} p(J_k) \prod_{k=0}^{K_\epsilon} h(\theta_k) \prod_{i=1}^{n} J_{s_i} v^{n-1} \exp\left\{-v \sum_{k=0}^{K_\epsilon} J_k\right\} \prod_{i=1}^{n} k(y_i|\theta_{s_i})$$

The Gibbs sampler

The full conditional of the Gibbs sampler are

1. The full conditional distribution of v is $\text{Ga}(n, \sum_{k=0}^{K_\epsilon} J_k)$.
2. The full conditional distribution of s_i is a discrete distribution with
$$p(s_i = k) \propto J_k p(y_i|\theta_k)$$
3. The full conditional distribution of θ_k has density proportional to
$$h(\theta_k) \prod_{\{i|s_i=k\}} k(y_i|\theta_{s_i})$$

4. The standard theory for NRMI priors can be used to define the following scheme. Let J' be the set of jumps with observations allocated them the

$$J'_k \sim \mathrm{Ga}(n_k, 1+v), \qquad J'_k > \epsilon$$

where n_k is the number of observation allocated to the k-th jump in J'.

Let $\tilde{J}$ be the jumps without allocated observations then the $\tilde{J}$ follow an ϵ-approximation with Lévy intensity

$$\exp\{-vJ\}\nu(J), \qquad J > \epsilon.$$

Series representations

We have looked at a few series representation but there are many.

Campbell et al. (2019) provide an overview of previous work. See also Nguyen et al. (2021) and Lee et al. (2023).

Slice sampling NRM mixtures (Griffin and Walker, 2011)

It's necessary to introduce latent variables $v_1, v_2, \ldots, v_n$ to fit these models using slice sampling and define

$$p(y, u, v|s) \propto v^{n-1} \prod_{i=1}^{n} I(u_i < J_{s_i}) \exp\left\{-v \sum_{k=1}^{\infty} J_k\right\} k(y_i|\theta_{s_i}).$$

which involves an infinite number of jumps.

Slice sampling NRM mixtures (Griffin and Walker, 2011)

It's necessary to introduce latent variables $v_1, v_2, \ldots, v_n$ to fit these models using slice sampling and define

$$p(y, u, v|s) \propto v^{n-1} \prod_{i=1}^{n} I(u_i < J_{s_i}) \exp\left\{-v \sum_{k=1}^{\infty} J_k\right\} k(y_i|\theta_{s_i}).$$

which involves an infinite number of jumps.

This is resolved by integrating out all jumps smaller than L where $L = \min\{u_i\}$ and define $J_1, J_2, \ldots, J_K > L$ then use

$$v^{n-1} \prod_{i=1}^{n} I(u_i < J_{s_i}) \exp\left\{-v \sum_{k=1}^{K} J_k\right\} \mathrm{E}\left[\exp\left\{-v \sum_{k=K+1}^{\infty} J_k\right\}\right] k(y_i|\theta_{s_i}).$$

Slice sampling NRM mixtures

The expectation is

$$\mathrm{E}\left[\exp\left\{-v\sum_{k=K+1}^{\infty}J_k\right\}\right]=\exp\left\{-\int_0^L(1-\exp\{-vx\})\,\nu(x)\,dx\right\}$$

due to definition of a Lévy intensity.

Gibbs sampling

The method can made efficient by jointly updating u and J using the results of James et al. (2009) in the following steps.

1. Let J' be the set of jumps with observations allocated to them and let $n_j = \#\{i' | s_i = j\}$ then the full conditional of $J_j \in J'$ is proportional to

$$J_j^{n_j} \exp\{-vJ_j\}\nu(J_j)$$

Gibbs sampling

The method can made efficient by jointly updating u and J using the results of James et al. (2009) in the following steps.

1. Let J' be the set of jumps with observations allocated to them and let $n_j = \#\{i' | s_i = j\}$ then the full conditional of $J_j \in J'$ is proportional to
$$J_j^{n_j} \exp\{-vJ_j\}\nu(J_j)$$
2. Simulate $u_1, u_2, \ldots, u_n$ where $u_i \sim \mathrm{U}(0, J_{s_i})$ and set $L = \min\{u_i\}$.

Gibbs sampling

The method can made efficient by jointly updating u and J using the results of James et al. (2009) in the following steps.

1. Let J' be the set of jumps with observations allocated to them and let $n_j = \#\{i' | s_i = j\}$ then the full conditional of $J_j \in J'$ is proportional to
$$J_j^{n_j} \exp\{-vJ_j\}\nu(J_j)$$
2. Simulate $u_1, u_2, \ldots, u_n$ where $u_i \sim \mathrm{U}(0, J_{s_i})$ and set $L = \min\{u_i\}$.
3. Simulate the jumps with no observations allocate to them from a Poisson process on (L, ∞) with intensity function $\exp\{-vx\}\nu(x)$.

Gibbs sampling for Dirichlet process mixtures

The method can made efficient by jointly updating u and J using the results of James et al. (2009) in the following steps.

Gibbs sampling for Dirichlet process mixtures

The method can made efficient by jointly updating u and J using the results of James et al. (2009) in the following steps.

1. Let J' be the set of jumps with observations allocated to them and let $n_j = \#\{i|s_i = j\}$ then the full conditional of $J_j \in J'$ is proportional to $J_j^{n_j-1}\exp\{-(1+v)J_j\}$, *i.e.* $\text{Ga}(n_j, 1+v)$.

Gibbs sampling for Dirichlet process mixtures

The method can made efficient by jointly updating u and J using the results of James et al. (2009) in the following steps.

1. Let J' be the set of jumps with observations allocated to them and let $n_j = \#\{i|s_i = j\}$ then the full conditional of $J_j \in J'$ is proportional to $J_j^{n_j-1}\exp\{-(1+v)J_j\}$, *i.e.* $\text{Ga}(n_j, 1+v)$.
2. Simulate $u_1, u_2, \ldots, u_n$ where $u_i \sim \text{U}(0, J_{s_i})$ and set $L = \min\{u_i\}$.

.

Gibbs sampling for Dirichlet process mixtures

The method can made efficient by jointly updating u and J using the results of James et al. (2009) in the following steps.

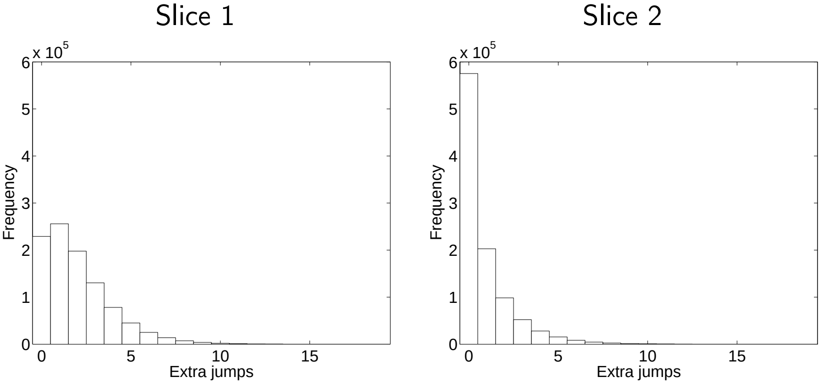
1. Let J' be the set of jumps with observations allocated to them and let $n_j = \#\{i|s_i = j\}$ then the full conditional of $J_j \in J'$ is proportional to $J_j^{n_j - 1} \exp\{-(1+v)J_j\}$, *i.e.* $\text{Ga}(n_j, 1+v)$.
2. Simulate $u_1, u_2, \ldots, u_n$ where $u_i \sim \text{U}(0, J_{s_i})$ and set $L = \min\{u_i\}$.
3. Simulate the jumps with no observations allocate to them from a Poisson process on (L, ∞) with intensity function $Mx^{-1} \exp\{-(1+v)x\}$.

Slice 2

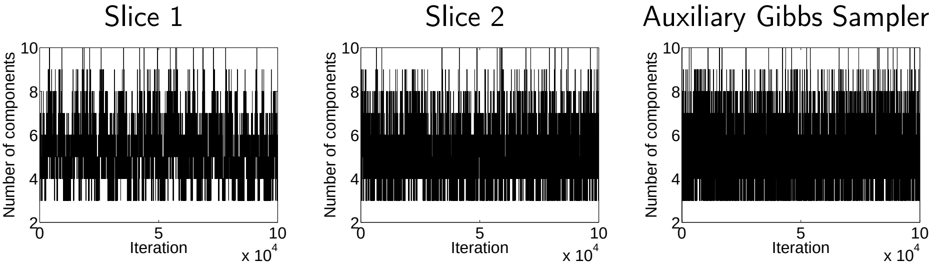
A slice sampler can also be defined with a single extra variable using the likelihood

$$v^{n-1}\prod_{i=1}^{n} I(u<\min\{J_{s_i}\})\frac{J_{s_i}}{\min\{J_{s_i}\}}\exp\left\{-v\sum_{k=1}^{\infty}J_k\right\}k(y_i|\theta_{s_i}).$$

Results - Galaxy Data / MDP



Comparisons - Galaxy / MDP



Results - Galaxy / NGG

	Number of Clusters		Number of empty clusters	
a	Mean	Standard Deviation	Mean	Standard Deviation
0.1	5.10	1.43	2.17	3.24
0.2	5.92	1.60	5.69	8.84
0.4	7.04	1.86	55.55	771.31

Andrieu, C. and G. O. Roberts (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics 37*, 697–725.

Campbell, T., J. H. Huggins, J. P. How, and T. Broderick (2019). Truncated random measures. *Bernoulli 25*, 1256–1288.

Dahl, D. B. and S. Newcomb (2022). Sequentially allocated merge-split samplers for conjugate bayesian nonparametric models. *Journal of Statistical Computation and Simulation 92*, 1487–1511.

Favaro, S. and Y.-W. Teh (2013). MCMC for normalized random measure mixture models. *Statistical Science 28*, 335–359.

Ferguson, T. S. and M. J. Klass (1972). A representation of independent increment processes without Gaussian components. *The Annals of Mathematical Statistics 43*, 1634–1643.

Griffin, J. E. (2019). Two part envelopes for rejection sampling of some completely random measures. *Statistics and Probability Letters 151*, 36–41.

Griffin, J. E. and S. G. Walker (2011). Posterior simulation for normalized random measure mixtures. *Journal of Computational and Graphical Statistics 20*, 241–259.

Ishwaran, H. and L. F. James (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association 96*, 161–173.

James, L. F., A. Lijoi, and I. Prünster (2009). Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics 36*, 76–97.

Kalli, M., J. E. Griffin, and S. G. Walker (2011). Slice sampling mixture models. *Statistics and Computing 21*, 93–105.

Lee, J., X. Miscouridou, and F. Caron (2023). A unified construction for series representations and finite approximations of completely random measures.

Neal, R. M. and S. Jian (2004). A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics 13*, 158–182.

Nguyen, T. D., J. Huggins, L. Masoero, L. Mackey, and T. Broderick (2021). Independent finite approximations for Bayesian nonparametric inference. arXiv:2009.10780.

Rodriguez, A. and D. B. Dunson (2011). Nonparametric bayesian models through probit stick-breaking processes. *Bayesian Analysis 6*, 145–178.

Rosinski, J. (2001). Series representations of l evy processes from the perspective of point processes. In O. Barndoff-Nielsen, S. Resnick, and T. Mikosch (Eds.), *Lévy processes: theory and applications*. Birkhaüser.

Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communcations in Statistics - Simulation and Computation 36*, 45–54.