An adaptive truncation method for inference in Bayesian nonparametric models

J. E. Griffin

Received: 10 December 2013 / Accepted: 25 September 2014 / Published online: 9 December 2014
© Springer Science+Business Media New York 2014

Abstract Many exact Markov chain Monte Carlo algorithms have been developed for posterior inference in Bayesian nonparametric models which involve infinite-dimensional priors. However, these methods are not generic and special methodology must be developed for different classes of prior or different models. Alternatively, the infinite-dimensional prior can be truncated and standard Markov chain Monte Carlo methods used for inference. However, the error in approximating the infinite-dimensional posterior can be hard to control for many models. This paper describes an adaptive truncation method which allows the level of the truncation to be decided by the algorithm and so can avoid large errors in approximating the posterior. A sequence of truncated priors is constructed which are sampled using Markov chain Monte Carlo methods embedded in a sequential Monte Carlo algorithm. Implementational details for infinite mixture models with stick-breaking priors and normalized random measures with independent increments priors are discussed. The methodology is illustrated on infinite mixture models, a semiparametric linear mixed model and a nonparametric time series model.

Keywords Sequential Monte Carlo · Dirichlet process · Poisson–Dirichlet process · Normalized random measures with independent increments · Truncation error · Stick-breaking priors

1 Introduction

The popularity of Bayesian nonparametric modelling has rapidly grown with computational power in a range of application areas such as epidemiology, biostatistics and economics. Bayesian nonparametric models have an infinite number of parameters over which a prior is placed and so allow model complexity to grow with sample size. In many models, the prior is placed on the space of probability measures. An overview of work in this area is given in Hjort et al. (2010). The infinite-dimensional prior does not allow the direct use of simulation-based methods for posterior inference or for the study of some properties of the distributions drawn from the prior (particularly, functionals such as the mean). The methods developed in this paper will concentrate on the former problem of posterior inference. Simulation-based methods require a finite-dimensional parameter space and there are two main approaches to working with such a space in the literature: marginalization and truncation.

The first approach marginalizes over the infinite-dimensional parameter and is particularly applicable to infinite mixture models such as the Dirichlet process mixture model. In these models, marginalization leads to Pólya urn scheme (PUS) representations of the priors which can be used to define efficient Markov chain Monte Carlo (MCMC) algorithms. Algorithms based on the PUS representation of Dirichlet process mixture models were reviewed in MacEachern (1998) and Neal (2000) and similar algorithms for normalized random measures mixture models were developed in Favaro and Teh (2013). These methods are limited by the unavailability of a suitable PUS for some priors.

The second method is truncation in which the infinite-dimensional prior is replaced by a finite-dimensional approximation. Approaches to the approximation of the Dirich-
let process were initially studied by Muliere and Tardella (1998) who showed how the error in total variation norm between their approximation and the infinite-dimensional prior can be chosen to be smaller than any particular value. Further results and alternative truncation methods for the Dirichlet process were developed by Ishwaran and Zarepour (2002). Ishwaran and James (2001) looked at truncating the wider-class of stick-breaking (SB) priors using total variation norm to measure truncation error and described a simple block Gibbs sampler for posterior inference. Argiento et al. (2010) consider the truncation of normalized generalized gamma process priors. More recently, slice sampling methods have been proposed for Dirichlet process (Walker 2007; Kalli et al. 2011), normalized random measures with independent increments (Griffin and Walker 2010) and σ-stable Poisson–Kingman (Favaro and Walker 2013) priors. In a slice sampler, the posterior distribution is augmented by auxiliary variables which lead to finite-dimensional distributions for all full conditionals of a Gibbs sampler and so represent a class of random truncation methods. Importantly, these methods sample exactly from the posterior distribution and so avoid truncation errors which are usually introduced by other truncation methods. However, like the marginalization methods, it is unclear how slice sampling methods can be applied generically to models with nonparametric priors.

The purpose of this paper is to develop a method for the choice of truncation level of nonparametric priors which adapts to the complexity of the data and which can be applied generically to nonparametric models. A sequence of finite-dimensional approximations to a nonparametric prior is constructed where the level of truncation is decreasing. A sequential Monte Carlo (SMC) approach (Del Moral et al. 2006) is used to sample from the corresponding sequence of posterior distributions. Some recent work in this area has emphasized the ability of these algorithms to adapt algorithmic tuning parameters to the form of the posterior distributions. Some theoretical aspects of these types of algorithm are discussed in Beskos et al. (2014).

The idea of using ESS as a measure of discrepancy will play a key role in adapting the tuning parameter (the level of truncation) in the methods developed in this paper. A truncation chosen when differences in the successive discrepancies become small will typically lead to small truncation errors. Importantly, special theory (over and above the definition of a sequence of approximating processes) is not needed to implement the method.

The paper is organized as follows. Section 2 is a review of some Bayesian nonparametric priors and truncation methods available for them. Section 3 describes the adaptive truncation algorithm for choosing the truncation level. Section 4 shows how this algorithm can be used to simulate from some popular nonparametric mixture models. Section 5 illustrates the use of these methods in infinite mixture models, and two non-standard nonparametric mixture models.

2 Finite truncation of infinite-dimensional priors

This paper will concentrate on a particular class of infinite-dimensional priors, random probability measures, which frequently arise in Bayesian nonparametric modelling. These probability measures are usually discrete with an infinite number of atoms and arise from constructions such as transformations of Lévy processes or SB priors. In these cases, the random probability measure $F$ can be expressed as

$$F = \sum_{j=1}^{\infty} p_j \delta_{\theta_j}$$

where $p = (p_1, p_2, \ldots)$ and $\theta = (\theta_1, \theta_2, \ldots)$ are sequences of random variables, $p_j > 0$ for $j = 1, 2, \ldots, \sum_{j=1}^{\infty} p_j = 1$, and $\delta_{\theta_j}$ is the Dirac delta measure which places measure 1 on $x$. Usually, it is further assumed that $p$ and $\theta$ are independent \textit{a priori}. This section will review the main constructions that fall within this class and truncation methods which have been proposed.

The Dirichlet process (Ferguson 1973) was originally defined as a normalized gamma process so that

$$F(B) = \frac{G(B)}{G(\Omega)}$$

where $B$ is a measurable set, $G$ is a Gamma process, i.e. a Lévy process with Lévy measure $\nu(J, \theta) = M J^{-1} \exp(-J) dJ H(d\theta)$ where $M > 0$ and $H$ is a probability measure (whose density, if it exists, is $h$) with support $\Omega$ and parameters $\psi$. Alternatively, we can write

$$F = \frac{\sum_{j=1}^{\infty} J_j \delta_{\theta_j}}{\sum_{j=1}^{\infty} J_j}$$

where $J = (J_1, J_2, \ldots)$ are the jumps of a Lévy process with Lévy density $M J^{-1} \exp(-J)$ and $\theta_j \sim H$. This construction can be naturally extended to an additive, increasing stochastic process (see e.g Nieto-Barajas et al. 2004).
A wide, tractable class of such priors can be defined by assuming that the jumps $J$ arise from a suitably-defined Lévy process, which is called the class of normalized random measures with independent increments (NRMIIs) (Regazzini et al. 2003). Examples, other than the Dirichlet process, include the normalized inverse Gaussian (NIG) process (Lijoi et al. 2005) and the normalized generalized gamma (NGG) process (Lijoi et al. 2005, 2007). Posterior inference for the general class is described in James et al. (2009) who assume a general formulation where $(J_1, \theta_1), (J_2, \theta_2), \ldots$ follows a Lévy process with Lévy density $v(J, \theta) = h(\theta | J) \eta(J)$. The process is called homogeneous if $h(\theta | J) = h(\theta)$ and inhomogeneous otherwise. The Dirichlet process, NIG process and NGG process are all homogeneous.

Truncated versions of these normalized processes can be constructed by truncating the Lévy process and then normalizing. This leads to a well-defined random probability measure. The simulation of finite-dimensional truncations of non-Gaussian Lévy processes is an active area of research.

An alternative truncation method uses the CPP approximation for NIG and Poisson–Dirichlet processes. Ishwaran and James (2001) study this truncation and show that it converges almost surely to the infinite-dimensional prior as $N \to \infty$. Favaro et al. (2012) derive an SB construction for the NIG process.

A truncated version of the prior with $N$ atoms can be defined by

$$F_N^\text{SB} = \sum_{j=1}^N p_j \delta_{\theta_j}$$

where $p_j = V_j \prod_{k<j} (1 - V_k)$ for $j = 1, \ldots, N$, $V_j \overset{\text{ind.}}{\sim} \text{Be}(a_j, b_j)$ for $j = 1, \ldots, N - 1$ and $V_N = 1$. This truncation is well-defined since $\sum_{j=1}^N p_j = 1$. Ishwaran and James (2001) study this truncation and show that it converges almost surely to the infinite-dimensional prior as $N \to \infty$. Muliere and Tardella (1998) define a similar truncation, which they term $\epsilon$-Dirichlet distribution, where $N$ is chosen to be the smallest value of $W$ for which $\sum_{j=1}^{W-1} p_j > 1 - \epsilon$ for some pre-specified $\epsilon$ with $F$ following a Dirichlet process. One drawback with this truncation is that the weights may no longer be stochastically ordered since $E[p_N] > E[p_{N-1}]$ if $b_{N-1} > a_{N-1}$. It can be shown that the weights are never stochastically ordered for a Poisson–Dirichlet process if $a > 0$. An alternative method of truncation normalizes the SB prior with $N$ atoms leading to the re-normalized stick-breaking (RSB) truncation

$$F_N^\text{RSB} = \sum_{j=1}^N \frac{p_j \delta_{\theta_j}}{1 - \prod_{j=1}^{N-1} (1 - V_j)}$$

where $p_j = V_j \prod_{k<j} (1 - V_k)$ for $j = 1, \ldots, N$, $V_j \overset{\text{ind.}}{\sim} \text{Be}(a_j, b_j)$ for $j = 1, \ldots, N$. Clearly, $p_1, \ldots, p_N$ are stochastically ordered and maintain that property of the infinite-dimensional SB prior.

An alternative form of truncation for the Dirichlet process (Ishwaran and Zarepour 2000; Neal 2000) uses finite-dimensional (FD) truncated distribution

$$F_N^\text{FD} = \sum_{j=1}^N \gamma_j \delta_{\theta_j}$$

where $\gamma_j \sim \text{Ga}(M/N, 1)$ (here $\text{Ga}(a, b)$ represents a gamma distribution with shape $a$ and mean $a/b$) and $\theta_j \overset{i.i.d.}{\sim} H$. 

The SB construction of the Dirichlet process dates from Sethuraman (1994) and expresses the weights in (1) as $p_j = V_j \prod_{k<j} (1 - V_k)$ where $V_j \overset{\text{i.i.d.}}{\sim} \text{Be}(1, M)$ which are independent of $\theta$ with $\theta_j \overset{i.i.d.}{\sim} H$. The use of more general forms of SB prior which assume that $V_j \overset{\text{ind.}}{\sim} \text{Be}(a_j, b_j)$ were popularized by Ishwaran and James (2001). They gave conditions on $a = (a_1, a_2, \ldots)$ and $b = (b_1, b_2, \ldots)$ for the process to be well-defined (so that $\sum_{j=1}^N p_j = 1$ almost surely). A nice feature of this construction is that the weights are stochastically ordered, i.e. $E[p_1] > E[p_2] > E[p_3] > \ldots$. The Poisson–Dirichlet process (Pitman and Yor 1997) arises when $a_j = 1 - a$ and $b_j = M + a_j$.
Ishwaran and Zarepour (2002) show that \( \int g(\theta) F_N^D(d\theta) \rightarrow \int g(\theta) F(d\theta) \) in distribution for any measureable function \( g \) and this truncation converges to the Dirichlet process.

The truncation methods described so far involve the choice of \( N \) for the Ferguson–Klass methods and truncated SB priors or the choice of \( L \) (the smallest jump) for the CPP approximation of the NRMII process. The success of any truncation method at approximating the infinite-dimensional prior or posterior will depend critically on this choice and so considerable effect has been devoted to the choice of these truncation parameters.

It is important to distinguish between two motivations for truncation. The first is studying the properties of the prior distribution (particularly, mean of functionals of the distribution) and the second is posterior inference using these priors. Initial work on truncation methods was motivated by the first consideration. Muliere and Tardella (1998) demonstrated that their \( \epsilon \)-Dirichlet distribution can be used to sample Dirichlet process functionals. The method uses samples generated by a PUS-based sampler for the posterior of a Dirichlet process mixture and exploits ideas from sampling prior functional of Dirichlet processes. The method was extended by Ishwaran and James (2002).

The second motivation for the truncation method is the approximation of the infinite-dimensional posterior by the posterior under a truncated prior. It is important to note that a “good” approximation of the prior will not necessarily lead to a “good” approximation of the posterior distribution. An example of a truncation which is designed to yield a “good” approximation to the posterior is the truncated SB process (Ishwaran and James 2001). They also consider defining an approximation to the posterior is the truncated SB process functionals. The method emphasises that the parameter space is growing with the truncation level \( k \). The truncated parameters \( \phi_{1:k} \) could include the first \( N_k \) atoms of a SB representation or the atoms in a CPP approximation with jump larger than \( L_k \). The prior distribution of the parameter for the \( k \)-th truncation will be denoted by \( \pi_k(\phi_{1:k}|\kappa) \).

I will also assume that it is relatively straightforward to sample values from \( \pi_k(\phi_k|\phi_{1:(k−1)}, \kappa) = \frac{\pi_k(\phi_1:k|\kappa)}{\pi_k(\phi_{1:(k−1)}|\kappa)} \) which is the distribution of \( \phi_k \) given \( \phi_{1:(k−1)} \) and \( \kappa \) under the \( k \)-th truncation. This is true for the truncations described in Sect. 2. A sequence of models can now be constructed by replacing the infinite-dimensional prior in (5) by the sequence of truncated priors leading to a \( k \)-th model of the form

\[
p(y_i|\phi_{1:k}, \lambda) = \sum p_j(\phi_{1:k}) k_j(y_i|\phi_{1:k}, \lambda), \quad \phi_{1:k} \sim \pi_k(\phi_{1:k}|\kappa), \quad \lambda \sim p(\lambda), \quad \kappa \sim p(\kappa)
\]

where \( k_j(y|\phi_{1:k}, \lambda) \) will typically depend on the level of truncation. This leads to a sequence of posterior distributions

\[
\pi_1(\phi_1, \lambda, \kappa|y), \quad \pi_2(\phi_{1:2}, \lambda, \kappa|y), \ldots \quad \text{for which the \( k \)-th posterior is}
\]

\[
\pi_k(\phi_{1:k}, \lambda, \kappa|y) \propto \prod_{i=1}^n p(y_i|\phi_{1:k}, \lambda) \pi_k(\phi_{1:k}|\kappa) p(\lambda) p(\kappa).
\]

This sequence will converge to the infinite-dimensional posterior (which will have the same mode of convergence as the prior).

Sequential Monte Carlo methods (see Doucet and Johansen 2011 for a review) can be used to efficiently simulate samples from the sequence of posterior distributions. The steps are outlined in Algorithm 1 which uses adaptive re-sampling (e.g. Del Moral et al. 2006) and MCMC updating for static models (Chopin 2002). The parameter \( b \) controls the amount
of re-sampling with smaller values of $b$ implying less re-sampling. The value $b = 0.7$ is chosen for the examples in this paper. The posterior expectation of a function $f$ under the $k$-th truncated posterior will be written as $E_k[f(\phi_{1:k}, \lambda, \kappa)]$ and can be unbiasedly estimated by

$$\sum_{j=1}^{S} w_k^{(j)} f(\phi_{1:k}^{(j)}, \lambda^{(j)}, \kappa^{(j)})$$

where $w_k^{(1)}, \ldots, w_k^{(S)}$ are the weights at the end of the $(k-1)$-th iteration of Algorithm 1. Many ways of re-weighting the particles in step 4(a) have been described in the SMC literature. Systematic resampling (Kitagawa 1996) is used in this paper but other methods are described in Doucet and Johansen (2011).

Simulate $S$ particles, $(\phi_{1}^{(1)}, \lambda^{(1)}, \kappa^{(1)}), \ldots, (\phi_{S}^{(S)}, \lambda^{(S)}, \kappa^{(S)})$ from the posterior distribution $\pi_1(\phi_1, \lambda, \kappa | y)$ and set $w_1^{(j)} = 1$ for $j = 1, 2, \ldots, S$.

At the $k$-th iteration,

1. Propose $\phi_{k+1}^{(j)}$ from the transition density $\pi_{k+1}(\phi_{k+1}^{(j)} | \phi_{k+1}^{(j)}, \lambda^{(j)}, \kappa^{(j)})$ for $j = 1, 2, \ldots, S$.
2. Update the weights $w_{k+1}^{(1)}, \ldots, w_{k+1}^{(S)}$ according to

$$w_{k+1}^{(j)} = \frac{w_{k}^{(j)} f(\phi_{k+1}^{(j)}, \lambda^{(j)}, \kappa^{(j)})}{\bar{w}_k^{(j)}}, \quad j = 1, \ldots, S$$

where

$$\bar{w}_k^{(j)} = \frac{\prod_{i=1}^{n} p(y_i | \phi_{i}^{(j)}, \lambda^{(j)}, \kappa^{(j)})}{\prod_{i=1}^{n} p(y_i | \phi_{i}^{(k)}, \lambda^{(k)}, \kappa^{(k)})}$$

3. Calculate the effective sample size $ESS_k = \frac{\sum_{j=1}^{S} w_{k+1}^{(j)}}{\sum_{j=1}^{S} w_{k+1}^{(j)^2}}$.
4. If $ESS_k < bS$,
   (a) Re-weight the particles $(\phi_{1}^{(1)}, \lambda^{(1)}, \kappa^{(1)}), \ldots, (\phi_{S}^{(S)}, \lambda^{(S)}, \kappa^{(S)})$ in proportion to the weights $w_{1}^{(1)}, \ldots, w_{S}^{(S)}$.
   (b) Set $w_{1}^{(j)} = 1$ for $j = 1, 2, \ldots, S$.
   (c) Update $(\phi_{1}^{(1)}, \lambda^{(1)}, \kappa^{(1)})$, $(\phi_{2}^{(1)}, \lambda^{(1)}, \kappa^{(1)})$, $(\phi_{3}^{(1)}, \lambda^{(1)}, \kappa^{(1)})$ using $m$ MCMC iterations with stationary distribution $\pi_{k+1}(\phi_{1:k+1}, \lambda, \kappa | y)$ for $j = 1, \ldots, S$.

Algorithm 1: The adaptive truncation algorithm

In practice, samples can only be drawn from a finite number of posteriors, i.e., $\pi_1(\phi_1, \lambda, \kappa | y), \pi_2(\phi_1, \lambda, \kappa | y), \ldots, \pi_R(\phi_1, \lambda, \kappa | y)$. The truncation is made adaptive by choosing the value of $R$ during the run of the algorithm using the output of the SMC algorithm. Intuitively, the posterior distributions $\pi_k(\phi_{1:k}, \lambda, \kappa | y)$ will become increasingly similar as $k$ increases since the data will tend to have less effect on the posterior of $\phi_k$ as $k$ increases. For example, the probability of an observation being allocated to the $k$-th cluster decreases as $k$ increases and the posterior distribution of $\phi_k$ becomes increasingly like its prior distribution. In other words, $\pi_{k+1}(\phi_{1:(k+1)}, \lambda, \kappa | y)$ becomes increasingly similar to $\pi_{k+1}(\phi_{1:(k+1)}, \lambda, \kappa | y)$ for $k$ increases.

The remaining issue is the decision of when to stop the SMC sampler. It is useful to define $\psi_{k+1}$ to be the sample of values of $\phi_{1:(k+1)}$, $\lambda$, $\kappa$ and $w_{k+1}$ at the end of the $k$-th iteration. It is assumed that a discrepancy $D(\psi_{k+1})$ between $\pi_k$ and $\pi_{k+1}$ for $j < k$ can be calculated using $\psi_k$ and $\psi_{k+1}$. The discrepancy should be positive with $D(\psi_{k+1}) = 0$ if $\pi_k$ and $\pi_{k+1}$ are the same and increasing as $\pi_k$ and $\pi_{k+1}$ become increasingly different. Specific examples of such discrepancies will be discussed at the end of this section. I define the stopping point $R$ to be the smallest $T$ for which

$$D(\psi_{k+1}) < \delta \quad \text{for} \quad k = T - m + 1, \ldots, T$$

where $\delta$ and $m$ are chosen by the user. The parameter $\delta$ is positive and taken to be small (and whose value is considered in Sect. 5) and $m$ is usually fairly small (the choice $m = 3$ worked well in the examples). This makes operational the idea of the sequence posteriors “settling down”. It seems sensible to assume that this “settling down” indicates that $\pi_R(\cdot | y)$ is close to the posterior distribution for the infinite dimensional model.

One useful measure of the discrepancy between these two distributions available from the SMC sampler is the ESS (Liu 2001), which is further investigated in this paper. This is defined as

$$ESS_{k+1} = \frac{\left(\sum_{j=1}^{S} w_{k+1}^{(j)}\right)^2}{\sum_{j=1}^{S} w_{k+1}^{(j)^2}}.$$
4 Adaptive truncation algorithms for mixture models with some specific priors

This section describes adaptive truncation methods for mixture models in (5) with specific forms of nonparametric prior.

MCMC methods typically introduce latent allocation variables \( s_1, \ldots, s_n \) and re-express the model as

\[
y_i | s_i \overset{ind}{\sim} k_i(y_i | \phi, \lambda), \quad p(s_i = j) = p_j(\phi)
\]

\( i = 1, \ldots, n, \quad j = 1, 2, \ldots. \)

This approach will be used in the MCMC samplers described in this section.

4.1 SB priors

4.1.1 RSB truncation

The adaptive truncation algorithm can be used with the RSB truncation of the SB prior by first choosing an initial truncation with \( N_1 \) atoms, i.e. \( F_{N_1}^{RSB} \). This leads to an initial set of parameters \( \phi_1 = (V_1, \ldots, V_{N_1}, \theta_1, \ldots, \theta_{N_1}) \). The \( k \)-th truncated prior in the algorithm is \( F_{N_k}^{RSB} \) in (4) where \( N_k = N_1 + k - 1 \). The extra parameters introduced in the \( k \)-th truncation are \( \phi_k = (V_{N_k}, \theta_{N_k}) \) for \( k > 1 \) and their prior distribution is

\[
\pi_k (\phi_k | \phi_{1:(k-1)}, \kappa) = \text{Be} \left( V_{N_k}, a_{N_k}, b_{N_k} \right) h \left( \theta_{N_k} | \psi \right)
\]

where \( \kappa = (a, b, \psi) \). The adaptive truncation algorithm in Algorithm 1 can now be run. The MCMC sampler for the \( k \)-th truncation introduces the latent variables \( s_1, \ldots, s_n \) described at the start of this section which leads to a joint prior distribution of \( s \) given by

\[
p(s|V) = \prod_{i=1}^{n} \frac{p_{s_i}}{1 - \prod_{j=1}^{N_k} (1 - V_j)}.
\]

The normalization constant in the denominator leads to non-standard full conditional distribution for \( V_1, \ldots, V_{N_k} \). The identity \( \sum_{i=0}^{\infty} d^i = \frac{1}{1-d} \) if \( d < 1 \) leads to a representation of the prior, which introduces latent variables \( z_1, \ldots, z_n \), and has density

\[
p(s, z|V) = \prod_{i=1}^{n} p_{s_i} \left( \prod_{j=1}^{N_k} (1 - V_j) \right)^{z_i} \]

where \( z_i \geq 0 \) for \( i = 1, \ldots, n \) (see e.g. Antoniano-Villalobos and Walker 2012). It follows that \( \sum p(s, z) = p(s) \) where the sum is taken over all possible values of \( z \).

The augmented posterior \( \pi_k (\phi_k, \lambda, \kappa, s, z|y) \) has full conditional distributions of standard form which allows a Gibbs sampler to be run. The steps of the MCMC algorithm for \( \pi_k (\phi_{1:k}, \lambda, \kappa, s, z|y) \) are as follows.

**Updating s**

The full conditional distribution of \( s_i \) is

\[
p(s_i = j) \propto p_j k_j(y_i | \theta), \quad j = 1, 2, \ldots, N_k.
\]

**Updating z**

The full conditional density of \( z_i \) is proportional to

\[
\left( \prod_{j=1}^{N_k} (1 - V_j) \right)^{z_i}, \quad z_i = 0, 1, 2, \ldots
\]

which is a geometric distribution with success probability

\[
1 - \prod_{j=1}^{N_k} (1 - V_j).
\]

**Updating V**

The full conditional distribution of \( V_j \) is \( \text{Be} \left( a_j^*, b_j^* \right) \) where \( a_j^* = a_j + \sum_{i=1}^{n} I(s_i = j) \), and \( b_j^* = b_j + \sum_{i=1}^{n} I(s_i > j) + \sum_{i=1}^{n} z_i \) for \( j = 1, 2, \ldots, N_k \).

**Updating \( \theta \)**

The full conditional density of \( \theta_j \) is proportional to

\[
h(\theta_j | \psi) \prod_{i|s_i=j} k_j(y_i | \theta).
\]

4.1.2 SB truncation

In a similar way to the the RSB truncation of the SB prior, an initial truncation of the infinite sum with \( N_1 \) atoms is chosen, i.e. \( F_{N_1}^{SB} \), which has initial parameters \( \phi_1 = (V_1, \ldots, V_{N_1-1}, \theta_1, \ldots, \theta_{N_1}) \). The \( k \)-th truncated prior is \( F_{N_k}^{SB} \) in (3) where, again, \( N_k = N_1 + k - 1 \). The extra parameters introduced in the \( k \)-th truncation are \( \phi_k = (V_{N_k-1}, \theta_{N_k}) \) for \( k > 1 \) and their prior distribution is
\[ \pi_k(\phi_k \mid \theta_{k-1}, \kappa) = \text{Be}(V_{N_k-1} \mid a_{N_k-1}, b_{N_k-1}) h(\theta_{N_k} \mid \psi) \]

where \( \kappa = (a, b, \psi) \). The MCMC sampler for the \( k \)-th truncation is the same as the one described in the previous subsection with the exception that the full conditional distribution of \( V_j \) is \( \text{Be}(a^*_j, b^*_j) \) where \( a^*_j = a_j + \sum_{i=1}^n I(s_i = j) \) and \( b^*_j = b_j + \sum_{i=1}^n I(s_i > j) \) for \( j = 1, 2, \ldots, N_k - 1 \). This leads to a Gibbs sampler which has the exact form of the blocked Gibbs sampler Ishwaran and James (2001) for a truncation value \( N_k \).

4.2 Lévy process-based models

4.2.1 CPP truncation

In the CPP truncation, a sequence of truncation points \( L_1, L_2, \ldots \) is selected. There are many ways to choose these points. For example, an increment size \( \xi \) could be chosen and the sequence generated using \( L_k = L_{k-1} \exp[-\xi] = L_1 \exp[-(k-1)\xi] \). Alternatively, a sequence which satisfies \( \xi(L_k) = \xi(L_{k-1}) + 1 \) would imply that, on average, one atom is added to the truncation at each iteration. The initial truncation is then \( F_{L_1}^{CPP} \) which has initial parameters \( \phi_1 = (J_1, \ldots, J_{K_{L_1}}, \theta_1, \ldots, \theta_{K_{L_1}}) \). The \( k \)-th truncated prior is \( F_{L_k}^{CPP} \) in (2) and so the extra parameters introduced in the \( k \)-th truncation are

\[ \phi_k = (J_{K_{L_{k-1}}+1}, \ldots, J_{K_{L_k}}, \theta_{K_{L_{k-1}}+1}, \ldots, \theta_{K_{L_k}}) \]

which are the jumps whose jump size is between \( L_{k-1} \) and \( L_k \). The distribution of \( \phi_k \) conditional on \( \phi_{1:k-1} \) and \( \kappa \) under \( \pi_k \) is a marked Poisson process with intensity \( \eta(x) \exp[-v x] \) on \( (L_k, L_{k-1}) \) (for the jumps \( J_j \)) and mark distribution \( H \) (for the locations \( \theta_j \)). The MCMC samplers use a similar approach to the slice sampler of Griffin and Walker (2010) by introducing allocation variables \( s_1, \ldots, s_n \) and a latent variable \( v \) in an augmented prior

\[ p(s_1, \ldots, s_n, v) = v^{n-1} \prod_{i=1}^n J_{s_i} \exp\left\{ -v \sum_{j=1}^{K_{L_k}} J_j \right\}. \]

This leads to the correct marginal distribution \( p(s_1, \ldots, s_n) = \prod_{i=1}^n \frac{J_{s_i}}{\sum_{j=1}^{K_{L_k}} J_j} \).

The steps of the MCMC algorithm for \( \pi_k(\phi_k, \lambda, \kappa, s, v \mid y) \) are as follows.

**Updating \( J \) and \( \theta \)**

The parameters \( J \) and \( \theta \) are divided into two parts: the jumps to which observations have been allocated which will be denoted \( J^* \) and \( \theta^* \) and the jumps to which no observation has been allocated \( J^\perp \) and \( \theta^\perp \). The full conditional density of \( J^* \) is proportional to

\[ \eta(J_j^*) J_j^* \exp\left\{ -v J_j^* \right\}, \quad J_j^* > L \]

where \( m_j \) is the number of observations allocated to the \( l \)-th jump and the full conditional density of \( \theta_j^* \) is proportional to

\[ h(\theta_j^*) \prod_{i:s_i=j} k_j(y_i \mid \theta_j^*). \]

The full conditional of \( J^\perp \) is a marked Poisson process with intensity \( \exp[-v x] \) on \( (L_k, \infty) \) and mark distribution \( H \). The sampled values are then \( J = (J^*, J^\perp) \) and \( \theta = (\theta^*, \theta^\perp) \).

**Updating \( s \)**

The full conditional distribution of \( s_i \) is

\[ p(s_i = j) \propto J_j k_j(y_i \mid \theta), \quad j = 1, 2, \ldots, K_{L_k}. \]

**Updating \( v \)**

The full conditional distribution of \( v \) is \( \text{Ga}(n, \sum_{j=1}^{L_k} J_j) \).

5 Examples

5.1 Mixture models

The Dirichlet process mixture model (Lo 1984) is the most widely used Bayesian nonparametric model with many MCMC algorithms having been proposed (see MacEachern 1998; Griffin and Holmes 2010 for reviews). Therefore, these models represent a natural benchmark for new computational methods for Bayesian nonparametric inference. The adaptive truncation algorithm is not expected to outperform current MCMC methods for these models (in fact, its main purpose is to define a generic method for inference in non-standard nonparametric models) but it is useful to look at its performance in this standard model. The infinite mixture models considered in this subsection used \( k_j(y_i \mid \phi) = N(y_i \mid \mu_j, \sigma_j^2) \) where \( N(x \mid \mu, \sigma^2) \) represents the density of a normally distributed random variable with mean \( \mu \) and variance \( \sigma^2 \). Initially, a Dirichlet process with mass parameter \( M \) and centring measure \( H \) with density \( h(\mu, \sigma^2) = N(\mu \mid \mu_0, \sigma_0^2) \text{Ga}(\sigma^{-2} \mid \alpha, \beta) \) was considered for the nonparametric prior. The mixture models were applied to the galaxy data, which was first introduced into the Bayesian nonparametric literature by Escobar and West (1995). The observations were divided by 10,000. The hyperparameters were chosen to be \( \mu_0 = \bar{y}, \sigma^2 = 10, \alpha = 3 \) and \( \beta = 0.1(\alpha - 1)\delta^2 \) where \( \bar{y} \) and \( \delta^2 \) are the sample mean and the sample variance of the observations, which were chosen for the purposes of illustration.
Initially, the problem of density estimation with $M = 1$ was considered. The results of 20 different runs of the adaptive truncation algorithm using the RSB truncation with $m = 3$, different numbers of particles $S$ and different values of $\epsilon$ are shown in Fig. 1. The approximations of the posterior mean density improved as $S$ increases with smaller variability in the estimates but the effect of $\epsilon$ seemed negligible. These patterns were confirmed by calculating the mean integrated squared error (MISE) which measures the discrepancy between the approximations from the adaptive truncation algorithm and the infinite-dimensional posterior for different combinations of $S$ and $\epsilon$. The MISE is defined for fixed $S$ and $\epsilon$ as

$$\text{MISE} = \frac{1}{B} \sum_{i=1}^{B} \int \left( f^{(i)}(x) - f^{GS}(x) \right)^2 dx$$

where $B$ is the number of runs of the algorithm, $f^{(i)}(x)$ is the posterior mean density calculated using the output from the $i$-th run of the algorithm and $f^{GS}(x)$ is a “gold-standard” estimate from the infinite dimensional posterior. The MISE will be small if the density from the infinite-dimensional posterior is well approximated across all runs of the algorithm. The slice sampler of Kalli et al. (2011) was run with a burn-in of 50,000 iterations with a subsequent run length of 5,000,000 iterations as the gold-standard estimate. The results for the MISE with different values of $S$ and $\epsilon$ are shown in Table 1.

| $\epsilon$    | $S = 1000$ | $S = 3000$ | $S = 10000$ |
|--------------|-----------|-----------|-------------|
| $10^{-3}$    | $3.32 \times 10^{-4}$ | $2.47 \times 10^{-4}$ | $1.10 \times 10^{-4}$ |
| $10^{-4}$    | $3.83 \times 10^{-4}$ | $2.48 \times 10^{-4}$ | $1.37 \times 10^{-4}$ |
| $10^{-5}$    | $4.17 \times 10^{-4}$ | $1.62 \times 10^{-4}$ | $1.22 \times 10^{-4}$ |
| $10^{-6}$    | $3.97 \times 10^{-4}$ | $2.20 \times 10^{-4}$ | $1.10 \times 10^{-4}$ |

Table 1: The mean integrated squared error (MISE) over 20 runs of the adaptive truncation algorithm with RSB truncation of a Dirichlet process with $M = 1$ using $m = 3$, and different values of $S$ and $\epsilon$ for the galaxy data.

There were only small differences between the results with different values of $\epsilon$ for fixed values of $S$ but the approximations improved as $S$ increased for fixed $\epsilon$ (as we would expect).

A more challenging problem is inference about the hyperparameter $M$ of the Dirichlet process. Many truncation results previously developed in the literature assume a fixed value of $M$ and so do not easily generalize to this more complicated inference problem. The adaptive truncation algorithm with RSB truncation was run using 10,000 particles, $m = 3$, and different values of $\epsilon$. The hyperparameter $M$ was given an exponential prior with mean 1. Some results for posterior inference about $M$, the stopping time $R$ and the computational time are given in Table 2. The “MCMC” results were calculated using the slice sampling algorithm described in Kalli et al. (2011) which was run for the same amount of computational time.

Fig. 1 The posterior mean density from 20 different runs of the adaptive truncation algorithm using the RSB truncation of a Dirichlet process with $M = 1$ using $m = 3$, and different values of $S$ and $\epsilon$ for the galaxy data.
The number of iterations as the GS approximation used to calculate MISE. The slice sampler generates samples from the infinite-dimensional prior and so allowed quantification of the truncation error for different values of $\epsilon$. The adaptive truncation algorithm gave estimates of $E[M|y]$ which are very similar to those from the infinite dimensional posterior for all values of $\epsilon$. Table 2 also shows the results using the FK truncation with 10,000 particles. These typically had a larger error (although, the error was still not particularly large). As we would expected the stopping time increases on average as $\epsilon$ decreases for both the RSB and FK truncations. This leads to clearly increasing running times for the FK truncation but the effect is not clear with the RSB truncation where running times are very similar. This is due to the structure of the algorithm where computational effort is divided between running the MCMC sampler for $\pi_1(\cdot|y)$ and sequentially proposing from the transition density $\pi_{k+1}(\phi_{k+1}|\phi_k, \kappa)$. Difference in the stopping time need not have a large effect on overall computational time if the transition density can be sampled quickly relative to the MCMC sampler. The results are consistent with this observation. The transitions in the RSB truncation involves sampling a single beta random variables whereas the transition in the FK truncation involves a numerical inversion to find the next value in the Ferguson–Klass representation. The level of truncation error suggests that the RSB truncation should be preferred to the FK truncation in this problem.

The Dirichlet process has weights which decay exponentially in the SB representation. Other specifications of the nonparametric prior lead to a slower decay of the weights. One such prior is the Poisson–Dirichlet process (Pitman and Yor 1997). The rate of decay of the weights decreases as $a$ increases and large $a$ is associated with very slow decay of weights. This is an important test case for truncation methods since the slow decay of the weights can lead to large truncation errors unless many atoms are included in the approximation. The adaptive truncation algorithm was tested on the infinite mixture model described at the start of section where the Poisson–Dirichlet process was used as

| $\epsilon$ | RSB | FK |
|---|---|---|
| $10^{-3}$ | 0.846 | 0.884 |
| $10^{-4}$ | 0.840 | 0.848 |
| $10^{-5}$ | 0.848 | 0.862 |
| $10^{-6}$ | 0.842 | 0.863 |

The parameter $M$ is assumed unknown and the algorithmic parameters were $m = 3, S = 10,000$ and different values of $\epsilon$ for the galaxy data. The reported values are averaged over the 20 runs with sample standard deviations shown in brackets.

### Table 2

| $\epsilon$ | RSB | FK |
|---|---|---|
| $10^{-3}$ | 0.193 | 0.219 |
| $10^{-4}$ | 0.203 | 0.198 |
| $10^{-5}$ | 0.196 | 0.196 |

The parameters $a$ and $M$ were assumed unknown and the algorithmic parameters were $m = 3, S = 10,000$ and different values of $\epsilon$ for the galaxy data. The reported values are averaged over the 20 runs with sample standard deviations shown in brackets.

### Table 3

| $\epsilon$ | $E[a|y]$ | $E[M|y]$ |
|---|---|---|
| $10^{-3}$ | 0.219 | 0.569 |
| $10^{-4}$ | 0.203 | 0.574 |
| $10^{-5}$ | 0.198 | 0.577 |
| $10^{-6}$ | 0.196 | 0.573 |

The nonparametric prior. The parameters of the process were considered unknown. The parameter $a$ was given a uniform prior on $(0, 1)$ and $M$ was given an exponential prior with mean 1. Some results from the adaptive truncation algorithm with RSB truncation run with $m = 3$ and 10,000 particles are given in Table 3. The “MCMC” results were calculated using the method described for the Dirichlet process mixture model. In this case, the value of $\epsilon$ had some impact on the quality of approximation. The truncation error in estimating the posterior mean of $a$ and $M$ decreased as $\epsilon$ decreased as we would expect. The estimates with $\epsilon = 10^{-3}$ are close to the MCMC results and become very close for $\epsilon = 10^{-5}$ and $\epsilon = 10^{-6}$ illustrating that the adaptive truncation algorithm can work well in this challenging example. The price to be paid for the increased accuracy is typically larger stopping times which increases from 62.2 for $\epsilon = 10^{-3}$ to 1603.3 for $\epsilon = 10^{-6}$ and much longer computational times.

### 5.2 A semiparametric linear mixed model

Linear mixed models are a popular way to model the heterogeneity of subjects with repeated measurements. It is
assumed that responses \( y_{it} \), \( i = 1, \ldots, n \), and \( t = 1, \ldots, T \) are observed for the \( i \)-th subject with \( (1 \times p) \)-dimensional vectors of regressors \( X_{it} \), \( i = 1, \ldots, n \), and \( (1 \times q) \)-dimensional vectors of regressors \( Z_{it} \), \( i = 1, \ldots, n \). The vectors \( X_{it} \) and \( Z_{it} \) may have some elements in common. The usual linear mixed effects model assumes that

\[
y_{it} = X_{it} \beta + Z_{it} \gamma_i + \epsilon_{it}, \quad i = 1, \ldots, n, \quad t = 1, \ldots, T \tag{8}
\]

where \( \beta \) is a \( (p \times 1) \)-dimensional vector of fixed effects and \( \gamma_i \) is a \( (q \times 1) \)-dimensional vector of random effects for the \( i \)-th subject. The model is usually made identifiable by assuming that \( E[\gamma_i] = 0 \) and \( E[\epsilon_{it}] = 0 \), which implies that \( E[y_{it} | X_{it}, Z_{it}] = X_{it} \beta \) and allows the regression effects \( \beta \) to be interpreted in the same way as in a linear regression model. Often parametric distributions are chosen for the errors and the random effects with \( \epsilon_{it} \sim N(0, \sigma^2_{\epsilon}) \) and \( \gamma_i \sim N(0, \Sigma_{\gamma}) \) being standard choices. However, in general, little is often known a priori about the distribution of the errors or the random effects and many authors have argued for a nonparametric approach.

Bayesian nonparametric inference in linear mixed models was initially considered by Bush and MacEachern (1996) and Kleinman and Ibrahim (1998) and subsequently developed by Ishwaran and Katakara (2002). These models assume that \( \gamma_i \) is given a Dirichlet process mixture prior but use a parametric distribution for the errors. The mean of a Dirichlet process is a random variable and so the condition that \( E[\gamma_i] = 0 \) is not imposed on the model. Typically, an alternative model is used where

\[
y_{it} = X_{it} \beta^* + Z_{it} \gamma_i^* + \epsilon_{it}, \quad i = 1, \ldots, n, \quad t = 1, \ldots, T \tag{9}
\]

and it is assumed that all elements of \( Z_{it} \) appear in \( X_{it} \) and that \( X_{it}^* \) is defined to be a design matrix containing the elements of \( X_{it} \) not shared with \( Z_{it} \). This removes the need for the identifiability constraint on the random effect since (8) implies that \( E[\gamma_i^*] \) can be non-zero. Li et al. (2011) discuss using post-processing of MCMC samples from (9) to make inference about the parameters in (8).

An alternative approach to inference in linear mixed models directly imposes location constraints on the nonparametric prior. Kottas and Gelfand (2001) and Hanson and Johnson (2002) constructed error distribution with median zero using mixtures of uniforms and Polya tree priors respectively. Tokdar (2006) constructed nonparametric priors whose realizations are zero mean distributions using the symmetrized Dirichlet process. I considered imposing constraints on the nonparametric priors for \( \epsilon_{it} \) and \( \gamma_i \) so that \( E[\epsilon_{it}] = 0 \) and \( E[\gamma_i] = 0 \) using the method of Yang et al. (2010). They assumed that \( \epsilon_{it} = \tilde{\epsilon}_{it} - E[\tilde{\epsilon}_{it}] \) and \( \gamma_i = \tilde{\gamma}_i - E[\tilde{\gamma}_i] \) where \( \tilde{\epsilon}_{it} \) and \( \tilde{\gamma}_i \) are given nonparametric priors without a mean constraint.

I assumed that \( q = 1 \) and used versions of the CCV model (Griffin 2010) as the nonparametric priors,

\[
p(\tilde{\epsilon}_{it}) = \sum_{j=1}^{\infty} p_j^\epsilon N(\mu_j^\epsilon, a_j^\epsilon \sigma^2_{\epsilon}) \quad \text{and} \quad p(\tilde{\gamma}_i) = \sum_{j=1}^{\infty} p_j^\gamma N(\mu_j^\gamma, a_j^\gamma \sigma^2_{\gamma})
\]

where the Dirichlet process prior for \( (p_1^\epsilon, \mu_1^\epsilon), (p_2^\epsilon, \mu_2^\epsilon), \ldots \) has mass parameter \( M_\epsilon \) and centring measure \( N(0, (1-a_\epsilon) \sigma^2_{\epsilon}) \). Similarly, the Dirichlet process prior for \( (p_1^\gamma, \mu_1^\gamma) \) has mass parameter \( M_\gamma \) and centring measure \( N(0, (1-a_\gamma) \sigma^2_{\gamma}) \). This allows \( a_\epsilon \) and \( a_\gamma \) to be interpreted as scale of \( \epsilon_{it} \) and \( \gamma_i \) respectively. The parameters \( a_\epsilon \) and \( a_\gamma \) are smoothness parameters with smaller values of indicating a rough density with potentially more modes. Yang et al. (2010) used a truncated version of the Dirichlet process prior to fit these types of models but do not develop any specific theory for choosing the number of atoms. I will consider using the adaptive truncation algorithm with RSB truncation to avoid choosing this value before running any algorithms. Details of the algorithm are given in Appendix 1.

The method will be illustrated on the “schoolgirl” data set taken from the “DPpackage” in R. The data are the heights...
of 20 children measured at ages from 6 to 10 inclusive and the height of their mothers which were divided into three groups (short, medium or tall). The data are shown in Fig. 2.

The groups were included using dummy variables and age was included as a regressor. The intercept for each schoolgirl was assumed to be a random effect. This leads to \( n = 20, T = 5, p = 4 \) and \( q = 1 \) with \( Z_{it1} = 1 \) for all \( i \) and \( t \). The nonparametric model was fitted to the data with the following hyperpriors:

- \( a_{\epsilon} \sim \text{Be}(1, 19) \), \( \sigma_{\epsilon}^2 \sim \text{FT}(1, 0.01) \), \( M_{\epsilon} \sim \text{Ga}(1, 1) \), \( a_{\gamma} \sim \text{Be}(1, 19) \), \( \sigma_{\gamma}^2 \sim \text{FT}(1, 1) \) and \( M_{\gamma} \sim \text{Ga}(1, 1) \), \( \beta \sim N(0, 10^6 I_4) \). The notation \( \text{FT}(v, A) \) denotes the folded \( t \)-distribution which has the density

\[
p(x) = \left(1 + \frac{x}{A}\right)^{-(v+1)/2}, \quad x > 0.
\]

This is a heavy-tailed distribution which was proposed by Gelman (2006) for variance parameters in hierarchical models. The adaptive truncation algorithm was run with 10,000 particles, \( \epsilon = 10^{-5} \) and \( m = 3 \). The initial MCMC run to simulate values from \( \pi_1(\phi_1, \lambda, \kappa | y) \) used a burn-in period of 5,000 iteration with a thinning of every fifth value. The densities of the observational error, \( f_{\epsilon} \), and the random effect, \( f_{\gamma} \), were summarized by their posterior means which are shown in Fig. 3. In both cases, the densities clearly deviated from normality. The posterior mean of \( f_{\epsilon} \) had a clear negative skewness and the posterior mean of \( f_{\gamma} \) had a positive skewness. The posterior densities of the group means are illustrated in Fig. 4. These show some evidence of differences between the group means with the largest differences between the tall group and the other two groups with a much less marked difference between the small and medium group means.

5.3 A nonparametric time series model

Antoniano-Villalobos and Walker (2012) described a method for Bayesian nonparametric inference in stationary time series models. Suppose that \( z_1, \ldots, z_T \) are a stationary time series, their model assumes that the transition probability is

\[
p(z_t | z_{t-1}) = \frac{\sum_{j=1}^{\infty} p_j \mathcal{N}\left(\left(\begin{array}{c} z_{t-1} \\ z_t \end{array}\right) \mid \left(\begin{array}{c} \mu_j \\ \mu_j \end{array}\right), \sigma^2_j \left(\begin{array}{cc} 1 & \rho_i \\ \rho_i & 1 \end{array}\right)\right)}{\sum_{j=1}^{\infty} p_j \mathcal{N}(z_{t-1} | \mu_j, \sigma^2_j)}
\]

and that the distribution of the initial value is

\[
p(z_1) = \sum_{j=1}^{\infty} p_j \mathcal{N}(z_{t-1} | \mu_j, \sigma^2_j).
\]
The stationary distribution is \( \sum_{j=1}^{\infty} p_j N(\zeta_j|\mu_j, \sigma^2_j) \) and the nonparametric specification of the transition density allows dependence to be flexibly modelled. Bayesian nonparametric inference in this model involves placing a prior on \( G_{z} = \sum_{j=1}^{\infty} p_j \delta_{\mu_j,\sigma^2_j,\rho_j} \) and Antoniano-Villalobos and Walker (2012) show that the prior has large support if \( G_z \) is given a Dirichlet process prior.

In practice, many observed data series are non-stationary. A simple model for a non-stationary time series \( y_1, \ldots, y_T \) has the form

\[
y_t = \alpha_t + \epsilon_t, \quad t = 1, \ldots, T
\]

where \( \alpha_t \) is a random walk component and \( \epsilon_t \) is a stationary process component. A flexible specification of this model would assume that \( \alpha_t \) follows a random walk whose increments are drawn from an unknown distribution and \( \epsilon_t \) follows the nonparametric model of Antoniano-Villalobos and Walker (2012). The process for \( \alpha_t \) is

\[
\alpha_t = \alpha_{t-1} + v_t, \quad p(v_t) = \sum_{j=1}^{\infty} p_j N(v_t|\mu_j, \sigma^2_j)
\]

where \( (p_j, \mu_j, \sigma^2_j) \) are given a Dirichlet process prior with mass parameter \( M_\alpha \) and centring measure \( N(0, (1 - M_\alpha)\sigma^2) \). The stationary process \( \epsilon_t \) is given a variation on the nonparametric prior described by Antoniano-Villalobos and Walker (2012) which assumes that \( \epsilon_t = \tilde{\epsilon}_t - E[\tilde{\epsilon}_t], \sigma^2 = \alpha_\epsilon \sigma^2_{\epsilon} \) and \( G_{\epsilon} \sim DP(M_\epsilon H_\epsilon) \) where \( H_\epsilon \) is a zero-mean normal distribution with variance \( (1 - \alpha_\epsilon)\sigma^2_{\epsilon} \). This ensures that the stationary distribution of \( \epsilon_t \) has zero expectation. Details of the algorithm are given in Appendix 2.

As an illustration, the model was applied to measurements of the annual flow of the Nile at Ashwan from 1871 to 1970, which is available from the “datasets” package in R. The data were standardized by subtracting the mean and dividing by the standard deviation and are plotted in Fig. 5. The graph shows clear evidence of non-stationary with a higher average level of flow in the initial years of the sample. The nonparametric model was fitted to the data with the following hyperpriors: \( a_\alpha \sim Be(1, 19), \sigma^2_\alpha \sim FT(1, 0.01), M_\alpha \sim Ga(1, 1), a_\epsilon \sim Be(1, 19), \sigma^2_\epsilon \sim FT(1, 1), M_\epsilon \sim Ga(1, 1) \) and \( \alpha_1 \sim N(0, 1) \). The values of \( \sigma^2_\alpha \) and \( \sigma^2_\epsilon \) were centred over values which allow all the variation in the data to be explained by one of the components, which represents a conservative choice. The adaptive truncation algorithm with RSB truncation was run using 10,000 particles, \( \epsilon = 10^{-5} \) and \( m = 3 \). The initial MCMC run to simulate values from \( \pi(\phi_1, \lambda, \kappa|y) \) used a burn-in period of 10,000 with a thinning of every fifth value. The posterior mean of the trend \( \alpha_t \) is plotted in Fig. 5. This clearly shows that the average flow of the Nile fell over the initial period of the data. The posterior mean stationary density of \( \epsilon_t \) and the posterior mean density of \( v_t \) are shown in Fig. 6. The posterior mean stationary density of \( \epsilon_t \) had a slight positive skew and the posterior mean density of \( v_t \) is heavy tailed with a pronounced spike at the mode.

The transition density of the stationary component \( \epsilon_t \) is shown in Fig. 7. There was clear evidence of a departure from the assumptions of an AR(1) process (which would be represented by diagonal bands with the same colour). The transition density was negatively skewed for values of \( \epsilon_{t-1} \) less than 0 whereas the density was roughly symmetric for \( \epsilon_{t-1} \) greater than 0. The conditional mean of \( \epsilon_t \) increased more quickly with \( \epsilon_{t-1} \) for positive \( \epsilon_{t-1} \) compared to negative \( \epsilon_{t-1} \).

6 Discussion

This paper describes a method for adaptively choosing the truncation point for posterior inference in nonparametric models. Application to the infinite mixture models showed that these methods can be effective for both density estimation and inference about hyperparameters of the nonparametric prior. The adaptive truncation method can be easily applied to non-standard mixture models, such as those fitted in Sects. 5.2 and 5.3, which cannot be fitted with the infinite-dimensional prior using MCMC methods.

Fig. 5 The data and the posterior median of \( \alpha_t \) (solid line) with point-wise 95% credible interval for the Nile flow data.
The methods developed in this article have relatively simple proposals in the SMC steps and only update global parameters if Step 4 occurs. This works well in the examples considered here but the current method has potential for further development which may be particularly important for problems where the nonparametric prior is defined on a high-dimensional space. These include variation on the general proposal mechanisms described by Del Moral et al. (2006) and the generic SMC² method (Chopin et al. 2013) which allow generalization of type of SMC methods (with MCMC steps) used in the adaptive truncation algorithm.

This paper has concentrated on inference in mixture models with nonparametric priors which is the most popular class of models in Bayesian nonparametric modelling. The adaptive truncation method is generic and can be applied to a much wider class of models. One increasingly important class of models are latent variable models with an infinite number of latent variables or processes. Examples include infinite factor models (Bhattacharya and Dunson 2011), infinite aggregation models (Kalli and Griffin 2014), and linear models with Lévy process priors (Polson and Scott 2012). Future work will consider the application of the adaptive truncation algorithm to these models.

Appendices

A. 1: A semiparametric linear mixed model

The mixture distributions in the model are approximated using the RSB truncation leading to the k-th truncated distributions

\[ p(\tilde{\epsilon}_{it}) = \sum_{j=1}^{N_k} p_j^\epsilon N(\tilde{\epsilon}_{it} | \mu_j^\epsilon, \sigma_j^2) \]

\[ p(\tilde{\gamma}_i) = \sum_{j=1}^{N_k} p_j^\gamma N(\tilde{\gamma}_i | \mu_j^\gamma, \sigma_j^2) \]

where \( p_j^\epsilon = V_j^\epsilon \Pi_{\epsilon,j}(1-V_j^\epsilon) \) and \( p_j^\gamma = V_j^\gamma \Pi_{\gamma,j}(1-V_j^\gamma) \)

with \( V_j^\epsilon \sim \text{Be}(1, M_j^\epsilon) \) and \( V_j^\gamma \sim \text{Be}(1, M_j^\gamma) \) for \( j = 1, \ldots, N_k \). The parameters in the initial truncation are \( \phi_1 = (V_{1:N_1}^\epsilon, \mu_{1:N_1}^\epsilon, \mu_{1:N_1}^\gamma) \) and the extra parameters added to form the k-th truncation are \( \phi_k = (V_{N_k}^\epsilon, \mu_{N_k}^\epsilon, V_{N_k}^\gamma, \mu_{N_k}^\gamma) \).

The algorithm introduces allocation variables \( s_{it}^\epsilon \) and \( s_{it}^\gamma \) for \( \epsilon_{it} \) and \( \gamma_i \) respectively with \( i = 1, \ldots, n \) and \( j = 1, \ldots, T \).
These are defined by
\[ p(s_{it}^\epsilon = j) = p_j^\epsilon \quad \text{and} \quad p(s_{it}^\gamma = j) = p_j^{\gamma^\epsilon}, \quad j = 1, \ldots, N_k. \]

It is useful to define
\[ f_k(a_\epsilon, \sigma_\epsilon^2, a_r, \sigma_r^2, \beta, \mu^\epsilon, V^\epsilon, \mu^\gamma, V^\gamma) \]
\[ = \exp \left\{ -\frac{1}{2} \left[ \frac{\sum_{i=1}^n \sum_{j=1}^r d_{ij}^2}{a_\epsilon \sigma_\epsilon^2} + \frac{\sum_{j=1}^r d_{ij}^2}{a_r \sigma_r^2} - \frac{\sum_{j=1}^r c_j^2}{d^2} \right] \right\} \times (a_\epsilon \sigma_\epsilon^2)^{-n/2} (a_r \sigma_r^2)^{-n/2} \times d^{-n/2} \]
\[ \text{where } d_{ij}^\epsilon = y_{it} - x_{it} \beta - \mu^\epsilon s_{ij}^\epsilon + \mu^\gamma s_{ij}^\gamma - \mu^\gamma, \]
\[ c_j = \frac{\sum_{i=1}^n d_{ij}^\epsilon}{a_\epsilon \sigma_\epsilon^2} + \frac{d_{ij}^\epsilon}{a_r \sigma_r^2}, \quad \text{and } d = \frac{\tau R}{\sigma^2} + \frac{1}{a_r \sigma_r^2} \]
and the values are calculated for the \( k \)-th truncation. Since \( s^\epsilon \) and \( s^\gamma \) are included in the sampling, values are proposed in Step 1 of the adaptive truncation algorithm. At the \( k \)-th iteration, the transition for the allocation \( s_{i}^\gamma \) is
\[ s_{i}^\gamma = \begin{cases} N_k \quad \text{with probability } 1 - r^\gamma, \\ s_{i,t}^\gamma \quad \text{otherwise} \end{cases} \]

and the transition for the allocation \( s_{i,t}^\epsilon \) is
\[ s_{i,t}^\epsilon = \begin{cases} N_k \quad \text{with probability } 1 - r^\epsilon, \\ s_{i,t}^\epsilon \quad \text{otherwise} \end{cases} \]
\[ \text{where } r^\gamma = \frac{1 - \Pi_{j=1}^{N_k} (1 - V_j^\gamma)}{1 - \Pi_{j=1}^{N_k} (1 - V_j^\gamma) + p_{N_k}} \quad \text{and } r^\epsilon = \frac{1 - \Pi_{j=1}^{N_k} (1 - V_j^\epsilon)}{1 - \Pi_{j=1}^{N_k} (1 - V_j^\epsilon) + p_{N_k}}. \]

In this case, the weights in Step 2 are updated using
\[ a_k^{(i)} = \frac{f_{k+1}(a_{\epsilon}^{(i)}, \sigma_{\epsilon}^{2(i)}, a_r^{(i)}, \sigma_r^{2(i)}, \beta, \mu^\epsilon(i), V^\epsilon(i), \mu^\gamma(i), V^\gamma(i))}{f_k(a_{\epsilon}^{(i)}, \sigma_{\epsilon}^{2(i)}, a_r^{(i)}, \sigma_r^{2(i)}, \beta, \mu^\epsilon(i), V^\epsilon(i), \mu^\gamma(i), V^\gamma(i))}. \]

The MCMC sampler updates many parameters using a variation of the adaptive Metropolis–Hastings random walk algorithm of Atchadé and Rosenthal (2005) which allows the proposal density to be updated at each iteration of the sampler. It works as follows for a generic parameter value \( \tau \).

Let \( q^{(i)}(\tau, \tau') = N(\tau' | \tau, \sigma^2(i)) \) be the proposal density for \( \tau \) at the \( i \)-th iteration which is random walk proposal with variance \( \sigma^2(i) \) and let \( a_i \) be the acceptance probability of a Metropolis–Hastings move with this proposal. The proposed value \( \tau' \) is accepted or rejected in the standard way for a Metropolis–Hastings step. The variance of the proposal is updated in the following way
\[ \log \sigma^2(i+1) = \rho \left( \log \sigma^2(i) + i^{-c} (a_i - \hat{a}) \right) \]
where \( 0.5 < c \leq 1 \) (the value \( c = 0.55 \) was used in the examples), \( \hat{a} \) is a target acceptance rate (the conservative choice \( \hat{a} = 0.3 \) was used in the examples) and
\[ \rho(x) = \begin{cases} -b & x < -b, \\ x - b & -b \leq x \leq b, \\ b & x > b \end{cases} \]

where \( b \) is taken to be very large. The average acceptance rate for the Metropolis–Hastings update of the parameter will converge to \( \hat{a} \) over the run of the sampler. The steps of the MCMC algorithm to sample from the posterior with the \( k \)-th truncation are described below.

**Updating \( \mu_j^\epsilon \)**

The full conditional density of \( \mu_j^\epsilon \) is proportional to
\[ f_k(a_\epsilon, \sigma_\epsilon^2, a_r, \sigma_r^2, \beta, \mu^\epsilon, V^\epsilon, \mu^\gamma, V^\gamma) \exp \left\{ -\frac{\mu_j^\epsilon^2}{(1 - a_\epsilon) \sigma_\epsilon^2} \right\} \]

and the parameter is updated using the adaptive Metropolis–Hastings random walk step.

**Updating \( V_j^\epsilon \)**

The full conditional density of \( V_j^\epsilon \) is proportional to
\[ f_k(a_\epsilon, \sigma_\epsilon^2, a_r, \sigma_r^2, \beta, \mu^\epsilon, V^\epsilon, \mu^\gamma, V^\gamma) \left( 1 - V_j^\epsilon \right)^{M_k-1} \]

and the parameter is updated using the adaptive Metropolis–Hastings random walk step after taking the transformation \( \log V_j^\epsilon - \log \left( 1 - V_j^\epsilon \right) \).

**Updating \( M_k \)**

The full conditional distribution of \( M_k \) is \( \text{Be}(a^*, b^*) \) where \( a^* = 1 + N_k \) and \( b^* = 1 - \sum_{j=1}^{N_k} \log \left( 1 - V_j^\epsilon \right) \).

**Updating \( a_\epsilon \)**

The full conditional density of \( a_\epsilon \) is proportional to
\[ f_k(a_\epsilon, \sigma_\epsilon^2, a_r, \sigma_r^2, \beta, \mu^\epsilon, V^\epsilon, \mu^\gamma, V^\gamma) \times \exp \left\{ -\frac{N_k}{(1 - a_\epsilon) \sigma_\epsilon^2} \right( 1 - a_\epsilon \right)^{18 - N_k/2} \]

and the parameter is updated using the adaptive Metropolis–Hastings random walk step after taking the transformation \( \log a_\epsilon - \log (1 - a_\epsilon) \).
Updating $\sigma_\epsilon^2$

The full conditional density of $\sigma_\epsilon^2$ is proportional to

$$f_k \left( a_\epsilon, \sigma_\epsilon^2, a_\gamma, \sigma_\gamma^2, \beta, \mu^\epsilon, V^\epsilon, \mu^\gamma, V^\gamma \right) \times \exp \left\{ \frac{-N_k}{2} \left( \frac{\sigma_\epsilon^2}{1-a_\epsilon} \right) \right\} \left( 1 + \sigma_\epsilon^2 \right)^{-1}$$

and the parameter is updated using the adaptive Metropolis–Hastings random walk step after taking the transformation $\log \sigma_\epsilon^2$.

Updating $\mu_j^\gamma$

The full conditional density of $\mu_j^\gamma$ is proportional to

$$f_k \left( a_\epsilon, \sigma_\epsilon^2, a_\gamma, \sigma_\gamma^2, \beta, \mu^\epsilon, V^\epsilon, \mu^\gamma, V^\gamma \right) \exp \left\{ -\frac{\mu_j^2}{(1-a_\gamma)\sigma_\gamma^2} \right\}$$

and the parameter is updated using the adaptive Metropolis–Hastings random walk step.

Updating $V_j^\gamma$

The full conditional density of $V_j^\gamma$ is proportional to

$$f_k \left( a_\epsilon, \sigma_\epsilon^2, a_\gamma, \sigma_\gamma^2, \beta, \mu^\epsilon, V^\epsilon, \mu^\gamma, V^\gamma \right) \left( 1 - V_j^\gamma \right)^{M_j-1}$$

and the parameter is updated using the adaptive Metropolis–Hastings random walk step after taking the transformation $\log V_j^\gamma - \log \left( 1 - V_j^\gamma \right)$.

Updating $M_j$

The full conditional distribution of $M_j$ is Be $(a^*, b^*)$ where $a^* = N_k$ and $b^* = 1 - \sum_{j=1}^{N_k} \log \left( 1 - V_j^\gamma \right)$.

Updating $a_\gamma$

The full conditional density of $a_\gamma$ is proportional to

$$f_k \left( a_\epsilon, \sigma_\epsilon^2, a_\gamma, \sigma_\gamma^2, \beta, \mu^\epsilon, V^\epsilon, \mu^\gamma, V^\gamma \right) \times \exp \left\{ -\frac{\sum_{j=1}^{N_j} \nu_j^2}{2(1-a_\gamma)\sigma_\gamma^2} \right\} \left( 1 - a_\gamma \right)^{-N_k/2}$$

and the parameter is updated using the adaptive Metropolis–Hastings random walk step after taking the transformation $\log a_\gamma - \log (1-a_\gamma)$.

Updating $\sigma_\gamma^2$

The full conditional density of $\sigma_\gamma^2$ is proportional to

$$f_k \left( a_\epsilon, \sigma_\epsilon^2, a_\gamma, \sigma_\gamma^2, \beta, \mu^\epsilon, V^\epsilon, \mu^\gamma, V^\gamma \right) \times \exp \left\{ -\frac{\nu_j^2}{2(1-a_\gamma)\sigma_\gamma^2} \right\} \left( 1 + \sigma_\gamma^2 \right)^{-1}$$

and the parameter is updated using the adaptive Metropolis–Hastings random walk step after taking the transformation $\log \sigma_\gamma^2$.

Updating $s^\epsilon$

The full conditional distribution of $s_i^\epsilon$ for $i = 1, \ldots, n$ and $t = 1, \ldots, T$ is

$$p(s_i^\epsilon = j) \propto \exp \left\{ -\frac{1}{2} \left[ \sum_{i=1}^{T} e_i^\epsilon \frac{e_i^\epsilon}{a_\epsilon \sigma_\epsilon^2} - r_i^2 \frac{2a_\epsilon^2}{d} \right] \right\} \text{, } j = 1, \ldots, N_k$$

where $e_i^\epsilon = y_{it} - X_i \beta - \mu_j + \tilde{\mu}^\epsilon$ and $r_i^\epsilon = \sum_{i=1}^{T} \frac{e_i^\epsilon}{a_\epsilon \sigma_\epsilon^2} + \frac{d_j^\epsilon}{a_\epsilon \sigma_\epsilon^2}$.

Updating $s^\gamma$

The full conditional distribution of $s_i^\gamma$ for $i = 1, \ldots, n$ is

$$p(s_i^\gamma = j) \propto \exp \left\{ -\frac{1}{2} \left[ \sum_{i=1}^{T} e_i^\gamma \frac{e_i^\gamma}{a_\gamma \sigma_\gamma^2} - r_i^2 \frac{2a_\gamma^2}{d} \right] \right\} \text{, } k = 1, \ldots, N_k$$

where $e_i^\gamma = \mu_j - \tilde{\mu}^\gamma$ and $r_i^\gamma = \sum_{i=1}^{T} \frac{e_i^\gamma}{a_\gamma \sigma_\gamma^2} + \frac{d_j^\gamma}{a_\gamma \sigma_\gamma^2}$.

Updating $\beta$

We sample $\gamma_i \sim N(c_i/d, 1/d)$ for $i = 1, \ldots, n$ which allows us to simulate $\beta$ as $\beta \sim N(\mu^*, \Sigma^*)$ where

$$\mu^* = \left( \frac{\sum_{i=1}^{T} X_i' X_i}{a_\epsilon \sigma_\epsilon^2} + 10^{-6} I_p \right)^{-1}$$

$$\times \left( \frac{\sum_{i=1}^{T} y_{it} - \gamma_i - \mu_j + \tilde{\mu}^\epsilon}{a_\epsilon \sigma_\epsilon^2} \right)$$

and

$$\Sigma^* = \left( \frac{\sum_{i=1}^{T} X_i' X_i}{a_\epsilon \sigma_\epsilon^2} + 10^{-6} I_p \right)^{-1}.$$
A. 2: A nonparametric time series model

The distribution of $\epsilon_i | \epsilon_{i-1}$ is approximated by the $k$-th truncated version

$$ p(\epsilon_i | \epsilon_{i-1}) = \frac{\sum_{j=1}^{N_k} p_j^e \mathcal{N} \left( \epsilon_{i-1} \mid \mu_j^e, \sigma_j^2 \right)}{\sum_{j=1}^{N_k} p_j^e \mathcal{N} \left( \epsilon_{i-1} \mid \mu_j^e, \sigma_j^2 \right)} $$

and distribution of the initial value is

$$ p(\epsilon_1) = \sum_{j=1}^{N_k} p_j^e \mathcal{N} \left( \epsilon_1 \mid \mu_j^e, \sigma_j^2 \right). $$

where $p_j^e = \frac{V_j^e \Pi_{j=1}^{N_k-1} (1-V_j^e)}{1-\Pi_{j=1}^{N_k} (1-V_j^e)}$ with $V_j^e \sim \text{Be}(1, M_k)$ for $j = 1, \ldots, N_k$, which is the RSB truncation. The initial parameters are $\phi_1 = \{ V_1^1, \mu_1^1, \sigma_1 \}$ and the extra parameters added to form the $k$-th truncation are $\phi_k = \{ V_k^N, \mu_k^N, \sigma_k \}$.

The nonparametric mixture model for $\nu_t$ is not constrained and so the algorithm uses the Pólya urn scheme representation to sample the parameters of this part of the model. Let $s_{2t}^q, \ldots, s_{2T}^q$ be the allocation variables for $\nu_1, \ldots, \nu_T$ respectively, $K_\alpha$ be the number of distinct values in the sample $s_{2t}^q, \ldots, s_{2T}^q$ and $\mu_1^q, \ldots, \mu_{K_\alpha}^q$ be the distinct values. It is useful to define

$$ f_k(\alpha, \sigma_\epsilon^2, \mu^e, p^e, \rho, \alpha) = \prod_{t=2}^T \left( \frac{\sum_{j=1}^{N_k} p_j^e (1-\rho_j^2)^{-1/2} \exp \left( -\frac{1}{2} \frac{b_{j, t} + b_{j, t-1}^2 - 2\rho_j b_{j, t} b_{j, t-1}}{a_j \sigma_j^2 (1-\rho_j)^2} \right)}{\sum_{j=1}^{N_k} p_j^e \exp \left( -\frac{1}{2} \frac{b_{j, t}^2}{a_j \sigma_j^2} \right)} \right)^{T/2} \sum_{j=1}^{N_k} p_j^e \exp \left( -\frac{1}{2} \frac{b_{j, t}^2}{a_j \sigma_j^2} \right) $$

where $b_{j, t} = y_t - \alpha_t - \mu_j^\epsilon + \mu^\epsilon$ for $t = 1, \ldots, T$ and $j = 1, \ldots, N_k$ with values of the parameter from the $k$-th truncation.

The MCMC sampler updates many parameters using a variation of the adaptive random walk algorithm of Atchadé and Rosenthal (2005) described in section Appendix 1.

Updating $\mu_j^\epsilon$

The full conditional density of $\mu_j^\epsilon$ is proportional to

$$ f_k(\alpha, \sigma_\epsilon^2, \mu^\epsilon, p^\epsilon, \rho, \alpha) \exp \left( -\frac{\mu_j^\epsilon^2}{(1-\alpha)\sigma_\epsilon^2} \right) $$

and the parameter is updated using the adaptive Metropolis–Hastings random walk step.

Updating $V_j^\epsilon$

The full conditional density of $V_j^\epsilon$ is proportional to

$$ f_k(\alpha, \sigma_\epsilon^2, \mu^\epsilon, p^\epsilon, \rho, \alpha) \left( 1 - V_j^\epsilon \right)^{M_k-1} $$

and the parameter is updated using the adaptive Metropolis–Hastings random walk step after taking the transformation $\log V_j^\epsilon - \log \left( 1 - V_j^\epsilon \right)$.

Updating $\rho_j$

The full conditional density of $\rho_j$ is proportional to

$$ f_k(\alpha, \sigma_\epsilon^2, \mu^\epsilon, p^\epsilon, \rho, \alpha) \exp \left( -\frac{\rho_j}{(1-\alpha)\sigma_\epsilon^2} \right) $$

and the parameter is updated using the adaptive Metropolis–Hastings random walk step after taking the transformation $\log \rho_j - \log (1 - \rho_j)$.

Updating $M_\epsilon$

The full conditional distribution of $M_\epsilon$ is $\text{Be}(a^*, b^*)$ where $a^* = 1 + N_k$ and $b^* = 1 - \sum_{j=1}^{N_k} \log \left( 1 - V_j^\epsilon \right)$.

Updating $a_e$

The full conditional density of $a_e$ is proportional to

$$ f_k(\alpha, \sigma_\epsilon^2, \mu^\epsilon, p^\epsilon, \rho, \alpha) \exp \left( -\frac{\sum_{j=1}^{N_k} \mu_j^2}{(1-a_e)\sigma_\epsilon^2} \right) \left( 1-a_e \right)^{18-N_k/2} $$

and the parameter is updated using the adaptive Metropolis–Hastings random walk step after taking the transformation $\log a_e - \log (1 - a_e)$.

Updating $\sigma_\epsilon^2$

The full conditional density of $\sigma_\epsilon^2$ is proportional to

$$ f_k(\alpha, \sigma_\epsilon^2, \mu^\epsilon, p^\epsilon, \rho, \alpha) \times \exp \left( -\frac{\sum_{j=1}^{N_k} \mu_j^2}{(1-a_e)\sigma_\epsilon^2} \right) \left( \sigma_\epsilon^2 \right)^{-N_k/2} \left( 1 + \sigma_\epsilon^2 \right)^{-1} $$

and the parameter is updated using the adaptive Metropolis–Hastings random walk step after taking the transformation $\log \sigma_\epsilon^2$. 
Updating $\alpha_1, \ldots, \alpha_T$

The full conditional density of $\alpha_1$ is proportional to

$$\frac{N_k}{\sum_{j=1}^{N_k} \rho_j^* (1 - \rho_j^2)^{-1/2}} \times \exp \left\{ -\frac{1}{2} \frac{b^2_{1,j} + b^2_{2,j} - 2 \rho_j b_{2,j} b_{1,j}}{a_\alpha \sigma_\alpha^2 (1 - \rho_j)^2} \right\} \exp \left\{ -\frac{1}{2} \frac{\alpha^2_1}{\sigma_\alpha^2} \right\}.$$

The full conditional density of $\alpha_t$ for $t = 2, \ldots, T - 1$ is proportional to

$$\prod_{k=t}^{T-1} \frac{N_k}{\sum_{j=1}^{N_k} \rho_j^* (1 - \rho_j^2)^{-1/2}} \times \exp \left\{ -\frac{1}{2} \frac{b^2_{1,j} + b^2_{2,j} - 2 \rho_j b_{2,j} b_{1,j}}{a_\alpha \sigma_\alpha^2 (1 - \rho_j)^2} \right\} \times \exp \left\{ -\frac{1}{2} \frac{\alpha^2_t - (\alpha_t - \alpha_{t-1} - \mu^*_j)^2}{\sigma_\alpha^2 a_\alpha} \right\}.$$

The full conditional density of $\alpha_T$ is proportional to

$$\frac{N_k}{\sum_{j=1}^{N_k} \rho_j^* (1 - \rho_j^2)^{-1/2}} \times \exp \left\{ -\frac{1}{2} \frac{b^2_{1,j} + b^2_{2,j} - 2 \rho_j b_{2,j} b_{1,j}}{a_\alpha \sigma_\alpha^2 (1 - \rho_j)^2} \right\} \times \exp \left\{ -\frac{1}{2} \frac{\alpha^2_T - (\alpha_T - \alpha_{T-1} - \mu^*_j)^2}{\sigma_\alpha^2 a_\alpha} \right\}.$$

Updating $s^*_t$

The full conditional distribution of $s^*_t$ is

$$p(s^*_t = j) \propto \left( \sum_{l=2, l \neq t}^T I(s^*_l = j) \right) a_\alpha^{-1/2} \times \exp \left\{ -\frac{1}{2} \frac{(\alpha_t - \alpha_{t-1} - \mu^*_j)^2}{a_\alpha \sigma_\alpha^2} \right\}, \quad j = 1, \ldots, K^-_\alpha$$

and

$$p(s^*_t = K^-_\alpha + 1) \propto M_\alpha \exp \left\{ -\frac{1}{2} \frac{(\alpha_t - \alpha_{t-1})^2}{\sigma_\alpha^2} \right\}.$$

where $K^-_\alpha$ is the number of distinct values without the $t$-th allocation and $\mu^*_1, \ldots, \mu^*_K^-_\alpha$ are the corresponding distinct values.

Updating $\mu_j^*$

The full conditional distribution of $\mu_j^*$ is $N(\mu^*, \sigma^* \cdot 2)$ where

$$\mu^* = \frac{\sum_{t=2}^T I(s^*_t = j) (\alpha_t - \alpha_{t-1})/a_\alpha}{\sum_{t=2}^T I(s^*_t = j)/a_\alpha + 1/(1 - a_\alpha)}$$

and

$$\sigma^* \cdot 2 = \frac{\sigma_\alpha^2}{\sum_{t=2}^T I(s^*_t = j)/a_\alpha + 1/(1 - a_\alpha)}.$$

Updating $a_\alpha$

The full conditional density of $a_\alpha$ is proportional to

$$a_\alpha^{-(T-1)/2} (1 - a_\alpha)^{18 - K^-_\alpha/2} \times \exp \left\{ -\frac{1}{2} \frac{\alpha^2_t - (\alpha_t - \alpha_{t-1} - \mu^*_j)^2}{a_\alpha \sigma_\alpha^2} \right\} \times \exp \left\{ -\frac{\sum_{j=1}^{K^-_\alpha} m^*_j^2}{(1 - a_\alpha) \sigma_\alpha^2} \right\}.$$}

and the parameter is updated using the adaptive Metropolis–Hastings random walk step after taking the transformation $\log a_\alpha < \log (1 - a_\alpha)$.

Updating $\sigma_\alpha^2$

The full conditional density of $\sigma_\alpha^2$ is proportional to

$$\sigma_\alpha^{2-(T-1+K^-_\alpha)/2} \times \exp \left\{ -\frac{1}{2} \frac{\alpha^2_t - (\alpha_t - \alpha_{t-1} - \mu^*_j)^2}{a_\alpha \sigma_\alpha^2} \right\} \times \exp \left\{ -\frac{\sum_{j=1}^{K^-_\alpha} m^*_j^2}{(1 - a_\alpha) \sigma_\alpha^2} \right\} \left(1 + 100 \sigma_\alpha^2 \right)^{-1}.$$}

and the parameter is updated using the adaptive Metropolis–Hastings random walk step after taking the transformation $\log \sigma_\alpha^2$.

Updating $M_\alpha$

The full conditional density of $M_\alpha$ is proportional to

$$\frac{\Gamma(M_\alpha)}{\Gamma(M_\alpha + T - 1)} M_\alpha^{K^-_\alpha}$$

and the parameter is updated using the adaptive Metropolis–Hastings random walk step after taking the transformation $\log M_\alpha$.
References

Al Labadi, L., Zarepour, M.: On asymptotic properties and almost sure approximation of the normalized inverse-Gaussian process. Bayesian Anal. 8, 553–568 (2013)

Al Labadi, L., Zarepour, M.: On simulations from the two-parameter Poisson-Dirichlet process and the normalized inverse-Gaussian process. Sankhya 76, 158–176 (2014)

Antoniano-Villalobos, I., Walker, S.G.: A nonparametric model for stationary time series. Technical report, University of Kent (2012)

Argiento, R., Guglielmi, A., Pievato, A.: Bayesian density estimation and model selection using nonparametric hierarchical mixtures. Comput. Stat. Data Anal. 54, 816–832 (2010)

Atchadé, Y.F., Rosenthal, J.S.: On adaptive Markov chain Monte Carlo algorithms. Bernoulli 11, 815–828 (2005)

Beskos, A., Jasra, A., Kantas, N., Thiery, A.: On the convergence of adaptive SMC methods. Technical report (2014)

Bhattacharya, A., Dunson, D.B.: Sparse Bayesian infinite factor models. Biometrika 98, 291–306 (2011)

Bush, C.A., MacEachern, S.N.: A semiparametric bayesian model for randomised block designs. Biometrika 83, 275–285 (1996)

Chopin, N.: A sequential particle filter for static models. Biometrika 89, 539–551 (2002)

Chopin, N., Jacob, P.E., Papaspiliopoulos, O.: SMC²: an efficient algorithm for sequential analysis of state space models. J. R. Stat. Soc. Ser. B 75, 397–426 (2013)

Cont, R., Tankov, P.: Financial Modelling with Jumps Processes. Chapman & Hall / CRC Press, London (2008)

Del Moral, P., Doucet, A., Jasra, A.: Sequential Monte Carlo samplers. J. R. Stat. Soc. Ser. B 68, 411–436 (2006)

Del Moral, P., Doucet, A., Jasra, A.: An adaptive sequential Monte Carlo method for approximate Bayesian computation. Stat. Comput. 22, 1009–1020 (2012)

Doucet, A., Johansen, A.M.: A tutorial on particle filtering and smoothing: fifteen years later. In: Crisan, D., Rozovsky, B. (eds.) The Oxford Handbook of Nonlinear Filtering. Oxford University Press, Oxford (2011)

Escobar, M.D., West, M.: Bayesian density-estimation and inference using mixtures. J. Am. Stat. Assoc. 90, 577–588 (1995)

Favaro, S., Lijoi, A., Prünster, I.: On the stick-breaking representation of normalized inverse Gaussian priors. Biometrika 99, 663–674 (2012)

Favaro, S., Teh, Y.W.: MCMC for normalized random measure mixture models. Stat. Sci. 28, 335–359 (2013)

Favaro, S., Walker, S.G.: Slice sampling σ-stable Poisson–Kingman mixture models. J. Comput. Gr. Stat. 22, 830–847 (2013)

Ferguson, T.S.: A Bayesian analysis of some nonparametric problems. Ann. Stat. 1, 209–230 (1973)

Ferguson, T.S., Klass, M.J.: A representation of independent increment processes without Gaussian components. Ann. Math. Stat. 43, 1634–1643 (1972)

Gelfand, A.E., Kottas, A.: A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models. J. Comput. Gr. Stat. 11, 289–305 (2002)

Gelman, A.: Prior distributions for variance parameters in hierarchical models. Bayesian Anal. 1, 515–533 (2006)

Griffin, J.E.: Default priors for density estimation with mixture models. Bayesian Anal. 5, 45–64 (2010)

Griffin, J.E., Holmes, C.C.: Computational issues arising in Bayesian nonparametric hierarchical models. In: Hjort, N.L., Holmes, C.C., Mueller, P., Walker, S.G. (eds.) Bayesian Nonparametrics, pp. 208–222. Cambridge University Press, Cambridge (2010)

Griffin, J.E., Walker, S.G.: Posterior simulation of normalized random measure mixtures. J. Comput. Gr. Stat. 20, 241–259 (2010)

Hanson, T., Johnson, W.O.: Modeling regression error with a mixture of Polya trees. J. Am. Stat. Assoc. 97, 1020–1033 (2002)

Hjort, N.L., Holmes, C.C., Mueller, P., Walker, S.G.: Bayesian Nonparametrics. Cambridge University Press, Cambridge (2010)

Ishwaran, H., James, L.: Gibbs sampling methods for stick-breaking priors. J. Am. Stat. Assoc. 96, 161–173 (2001)

Ishwaran, H., James, L.J.: Approximate Dirichlet process computing in finite normal mixtures: smoothing and prior information. J. Comput. Gr. Stat. 11, 508–532 (2002)

Ishwaran, H., Takahara, G.: Independent and identically distributed Monte Carlo algorithms for semiparametric linear mixed models. J. Am. Stat. Assoc. 97, 1154–1166 (2002)

Ishwaran, H., Zarepour, M.: Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. Biometrika 87, 371–390 (2000)

Ishwaran, H., Zarepour, M.: Exact and approximate sum representations for the Dirichlet process. Can. J. Stat. 30, 269–283 (2002)

James, L.F., Lijoi, A., Prünster, I.: Posterior analysis for normalized random measures with independent increments. Scand. J. Stat. 36, 76–97 (2009)

Jasra, A., Stephens, D.A., Doucet, A., Tsagaris, T.: Inference for Lévy-driven stochastic volatility models via adaptive sequential Monte Carlo. Scand. J. Stat. 38, 1–22 (2011)

Kalli, M., Griffin, J.E.: Flexible modelling of dependence in volatility processes. J. Bus. Econ. Stat. (2014) (in press)

Kalli, M., Griffin, J.E., Walker, S.G.: Slice sampling mixture models. Stat. Comput. 21, 93–105 (2011)

Kitagawa, G.: Monte Carlo filter and smoother for non-Gaussian non-linear state space models. J. Comput. Gr. Stat. 5, 1–25 (1996)

Kleinman, K.P., Ibrahim, J.G.: A semiparametric Bayesian approach to the random effect model. Biometrics 54, 921–938 (1998)

Kottas, A., Gelfand, A.E.: Bayesian semiparametric median regression modeling. J. Am. Stat. Assoc. 96, 1458–1468 (2001)

Li, Y., Müller, P., Lin, X.: Center-adjusted inference for a nonparametric Bayesian random effect distribution. Stat Sinica 21, 1201–1223 (2011)

Lijoi, A., Mena, R., Prünster, I.: Bayesian nonparametric analysis for a generalized Dirichlet process prior. Stat. Inference Stoch. Process. 8, 283–309 (2005)

Lijoi, A., Mena, R.H., Prünster, I.: Hierarchical mixture modeling with normalized inverse-Gaussian priors. J. Am. Stat. Assoc. 100, 1278–1291 (2005)

Lijoi, A., Mena, R.H., Prünster, I.: Controlling the reinforcement in Bayesian non-parametric mixture models. J. R. Stat. Soc. Ser. B 69, 715–740 (2007)

Liu, J.S.: Monte Carlo Strategies in Scientific Computing. Springer-Verlag, New York (2001)

Lo, A.Y.: On a class of Bayesian nonparametric estimates: I. Density estimates. Ann. Stat. 12, 351–357 (1984)

MacEachern, S.N.: Computational methods for mixture of Dirichlet process models. In: Dey, D., Mueller, P., Sinha, D. (eds.) Practical Nonparametric and Semiparametric Bayesian Statistics, pp. 23–44. Springer-Verlag, New York (1998)

Muliere, P., Tardella, L.: Approximating distributions of random functions of Ferguson–Dirichlet priors. Can. J. Stat. 26, 283–297 (1998)

Neal, R.M.: Markov chain sampling methods for Dirichlet process mixture models. J. Comput. Gr. Stat. 9, 249–265 (2000)

Nieto-Barajas, L.E., Prünster, I., Walker, S.G.: Normalized random measures driven by increasing additive processes. Ann. Stat. 32, 2343–2360 (2004)

Pitman, J., Yor, M.: The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. Ann. Probab. 25, 855–900 (1997)
Polson, N.G., Scott, J.G.: Local shrinkage rules, Lévy processes, and regularized regression. J. R. Stat. Soc. Ser. B 74, 287–311 (2012)
Regazzini, E., Lijoi, A., Prünster, I.: Distributional results for means of normalized random measures with independent increments. Ann. Stat. 31, 560–585 (2003)
Schäfer, C., Chopin, N.: Sequential Monte Carlo on large binary sampling spaces. Stat. Comput. 23, 163–184 (2013)
Sethuraman, J.: A constructive definition of Dirichlet priors. Stat. Sinica 4(2), 639–650 (1994)
Tokdar, S.: Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. Sankhya 68, 90–110 (2006)
Walker, S.G.: Sampling the Dirichlet mixture model with slices. Commun. Stat. Simul. Comput. 36, 45–54 (2007)
Yang, M., Dunson, D.B., Baird, D.: Semiparametric Bayes hierarchical models with mean and variance constraints. Comput. Stat. Data Anal. 54, 2172–2186 (2010)