SMC samplers for finite and infinite mixture models

María Lomelí
University of Cambridge

SMC workshop
September 1, 2017

Table of Contents I

Motivation

Motivation: examples of compositional models with a BNP component

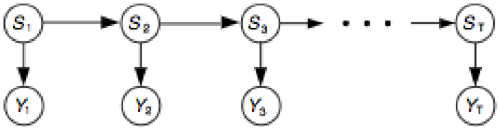


Figure: Infinite Hidden Markov Model[Beal et al., 2002]

$$
\begin{array}{rcl}
\beta|\gamma & \sim & \mathrm{Stick}(\cdot|\gamma) \\
\pi_k|\alpha,\beta & \sim & \mathrm{DP}(\cdot|\alpha,\beta) \\
\theta_k|H & \sim & H(\cdot) \\
s_t|s_{t-1},(\pi_k)_{k=1}^{\infty} & \sim & \pi_{s_{t-1}}(\cdot) \\
y_t|s_t,(\theta_k)_{k=1}^{\infty} & \sim & p(\cdot|\theta_{s_t})
\end{array}
$$

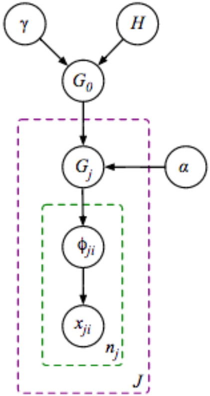


Figure: Hierarchical Dirichlet process [Teh et al., 2004]

[Pictures borrowed from Zoubin's UAI tutorial]

Mixture models

Mixture models

Let $\{Y_i\}_{i=1}^n$ be our data. A mixture model is an example of a latent variable model which has a single discrete latent variable per observation

$$
\begin{aligned}
X_i &\sim \mathrm{Categorical}(\underline{\pi}) \\
Y_i \mid X_i &\sim f(\cdot \mid \theta_{x_i}).
\end{aligned}
$$

Under the discrete distribution

$$
P(X_i = j) = \pi_j, \qquad \pi_j \geq 0, \qquad \sum_{j=1}^m \pi_j = 1
$$

and

$$
\begin{aligned}
P(Y_i \in \mathrm{d}y_i) &= \sum_{j=1}^m P(Y_i \in \mathrm{d}y_i \mid X_i) P(X_i = j) \\
&= \sum_{j=1}^m \pi_m F(\mathrm{d}y_i \mid \theta_j).
\end{aligned}
$$

In order to be fully Bayesian, a prior distribution for all unknown quantities should be incorporated

$$\begin{aligned} M &\sim \mathcal{Q} \\ \underline{\pi} \mid M = m &\sim \mathcal{P}_m \\ X_1, \cdots, X_n \mid \underline{\pi} &\overset{\text{i.i.d.}}{\sim} \text{Categorical}(\underline{\pi}) \\ Y_i \mid X_i &\sim f(\cdot \mid \theta_{X_i}). \end{aligned}$$

One option is to choose $\mathcal{Q}$ with support on $\mathbb{N}$.

In order to be fully Bayesian, a prior distribution for all unknown quantities should be incorporated

$$
\begin{aligned}
M &\sim \mathcal{Q} \\
\underline{\pi} \mid M = m &\sim \text{Symmetric Dirichlet}(\gamma) \\
X_1, \cdots, X_n \mid \underline{\pi} &\overset{\text{i.i.d.}}{\sim} \text{Categorical}(\underline{\pi}) \\
Y_i \mid X_i &\sim f(\cdot \mid \theta_{X_i}).
\end{aligned}
$$

One option is to choose $\mathcal{Q}$ with support on $\mathbb{N}$,, for instance

$$
\mathcal{Q}\,(M = m) = \frac{\eta(1-\eta)_{m-1\uparrow}}{m!}, \quad m \in \mathbb{N}, \quad \eta \in (0,1).
$$

Chinese Restaurant process as a limit

Let us assume there is an infinite total number of components.

Set $\gamma = \frac{\theta}{m}$

$$p(n_1, \cdots, n_k \mid M = m) = \frac{m! m^{-k}}{(m-k)!} \frac{\theta^k \Gamma(\theta)}{\Gamma(\theta + n)} \prod_{\{\ell : n_\ell > 0\}} \frac{\Gamma(n_\ell + \frac{\theta}{m})}{\Gamma(\frac{\theta}{m} + 1)}$$

Let $m \to \infty$

$$p(n_1, \cdots, n_k) = \frac{\theta^k \Gamma(\theta)}{\Gamma(\theta + n)} \prod_{\ell=1}^{k} \Gamma(n_\ell).$$

We have just derived the finite dimensional distribution of a Chinese restaurant process. [Aldous, 1985]

Infinite mixture models

Infinite mixture models

An infinite mixture model is a mixture model with potentially infinitely many mixture components.

$$
\begin{aligned}
G &\sim \text{Random probability measure (RPM)} \\
X_i \mid P &\sim P \\
Y_i \mid X_i &\sim F_{X_i}.
\end{aligned}
$$

[Lo, 1984, Rasmussen, 2000] choose G to be a Dirichlet process.

Random Probability measures

Any discrete distribution $G : \mathcal{B}(\mathbb{X}) \to [0,1]$ on a measurable space $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ can be represented as

$$G(B) = \sum_{i=1}^{\infty} p_i \delta_{z_i}, \quad B \in \mathcal{B}(\mathbb{X}), \quad \sum_{i=1}^{\infty} p_i = 1.$$

Make the weights $(P_i)_{i \in \mathbb{N}}$ and locations $(Z_i)_{i \in \mathbb{N}}$ random and you obtain that G is a **random probability measure**.

[Laha and Rohatgi, 1979, Kingman, 1975]

The Dirichlet and Pitman–Yor Processes as a Random Probability measure

Example 1: Dirichlet process (DP). Let $(V_i)_{i\in\mathbb{N}} \overset{\text{i.i.d}}{\sim} \text{Beta}(1,\theta)$ and $(Z_i)_{i\in\mathbb{N}} \overset{\text{i.i.d.}}{\sim} H_0$ independent of $(V_i)_{i\in\mathbb{N}}$. The stick breaking construction says

$$
\begin{aligned}
P_1 &= V_1 \\
P_i &= V_i \prod_{j<i}(1-V_j) \quad \forall i \geq 2
\end{aligned}
$$

Example 2: Pitman–Yor process (PY). Let $(V_i)_{i\in\mathbb{N}} \overset{\text{ind}}{\sim} \text{Beta}(1-\sigma,\theta+i\sigma)$ and $(Z_i)_{i\in\mathbb{N}} \overset{\text{i.i.d.}}{\sim} H_0$ independent of $(V_i)_{i\in\mathbb{N}}$. The stick breaking construction says

$$
\begin{aligned}
P_1 &= V_1 \\
P_i &= V_i \prod_{j<i}(1-V_j) \quad \forall i \geq 2
\end{aligned}
$$

[Sethuraman, 1994, Pitman and Yor, 1997]

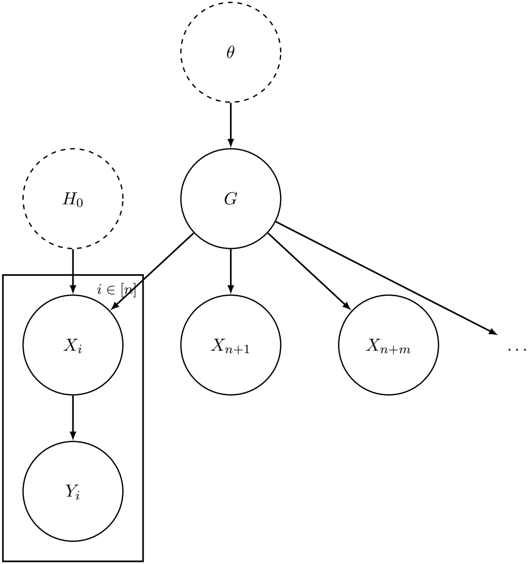


Figure: Intractable graphical model of a Dirichlet process mixture model

From Intractable to tractable representations

Clustering as a partition of the data

Partition of $[n] := \{1, \cdots, n\}$**,** $n \in \mathbb{N}$ **.** A partition $\Pi_n = \left\{A_1, \cdots, A_{|\Pi_n|}\right\}$ of the first n integers set [n], $n \in \mathbb{N}$ is a finite collection of $|\Pi_n|$ non-empty, non-overlapping and exhaustive subsets of $[n]$ called blocks and denoted by $A_j, j = 1, \cdots, |\Pi_n|$, i.e.

1. $\varnothing \subset A_j \subseteq [n], \forall j = 1, \cdots, |\Pi_n|$.
2. $A_i \cap A_j = \varnothing,\ \forall i, j \in [n], i \neq j$.
3. $\bigcup_{j=1}^{|\Pi_n|} A_j = [n]$.

$|\Pi_n|$ is the cardinality or number of blocks of the partition.

A **Chinese restaurant process** is a distribution over partitions of $\mathbb{N}$ whose finite dimensional distributions, called Exchangeable random probability functions (EPPF), have a particular form.

Family of Gibbs-type random partitions

An exchangeable random partition Π of the set of natural numbers $\mathbb{N}$ is said to be of *Gibbs form* with parameter $\sigma \in [-\infty, 1)$ if the EPPF of Π_n, $n \in \mathbb{N}$ satisfies

$$p(\Pi_n = \pi) = V_{n,k} \prod_{A \in \pi} \frac{\Gamma(|A| - \sigma)}{\Gamma(1 - \sigma)}$$

$\forall\, k \in \{1, \cdots, n\}$. It depends only on n: the number of observations, k: the number of blocks and the sizes of each block in the partition.

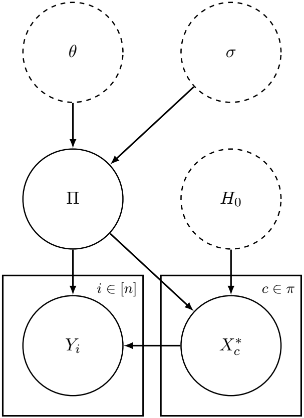


Figure: Tractable graphical model of a two parameter Chinese restaurant mixture model

First SMC sampler

Urn sequential construction

The predictive distribution of Gibbs type priors with parameter $\sigma \in (-\infty, 1)$ is given by

$$\Pr\left(X_{n+1} \in \cdot \mid X_1 \cdots, X_n\right) = \frac{V_{n+1,k+1}}{V_{n,k}} H_0(\cdot) + \frac{V_{n+1,k}}{V_{n,k}} \sum_{\ell=1}^{k} (n_\ell - \sigma) \delta_{X_\ell^*}(\cdot).$$

Finite number of total components case, $\sigma < 0$, and

$$V_{n,k} = \sum_{m=1}^{\infty} |\sigma|^k \frac{m\Gamma(m)}{\Gamma(m-k+1)} \frac{\Gamma(m|\sigma|)}{\Gamma(m|\sigma|+n)} \mathcal{Q}(m)$$

Infinite number of total components case, $\sigma \in (0,1)$, and

$$V_{n,k} = \int_{\mathbb{R}^+} \frac{\sigma^k}{\Gamma(n - \sigma k)} \left(t^{-\sigma}\right)^k \int_0^1 p^{n-\sigma k-1} f_\sigma\left((1-p)t\right) dph(t)dt$$

SMC proposal and incremental weight

$$\Pr\left(\text{i joins cluster c'} \mid \Pi^{\ell}_{i-1}, \mathbf{y}_{1:i-1}\right)$$
$$\propto \left\{ \begin{array}{cc} \frac{V_{n+\mathbf{1},k}}{V_{n,k}} f\left(y_i) \mid \{y_j\}_{j\in c'}\right) & \text{if } c' \in \Pi^{\ell}_{i-1} \\ \frac{V_{n+\mathbf{1},k+\mathbf{1}}}{V_{n,k}} f\left(y_i\right) & \text{o.w.} \end{array} \right\}$$

where

$$f\left(y_i\right) = \int f(y^i \mid \theta) H_0(d\theta)$$
$$f\left(y_i) \mid \{y_j\}_{j\in c}\right) = \int f\left(y^i \mid \theta\right) f\left(\theta \mid \{y_j\}_{j\in c}\right) \mathrm{d}\theta,$$

and the incremental weight is

$$p\left(y_i \mid \Pi^{\ell}_{i}, \mathbf{y}_{1:i-1}\right)$$
$$= \frac{V_{n+1,k+1}}{V_{n,k}} f(y^i) + \sum_{c\in\Pi^{\ell}_{i}} \frac{V_{n+1,k}}{V_{n,k}} f(y^i \mid \{y_j\}_{j\in c}).$$

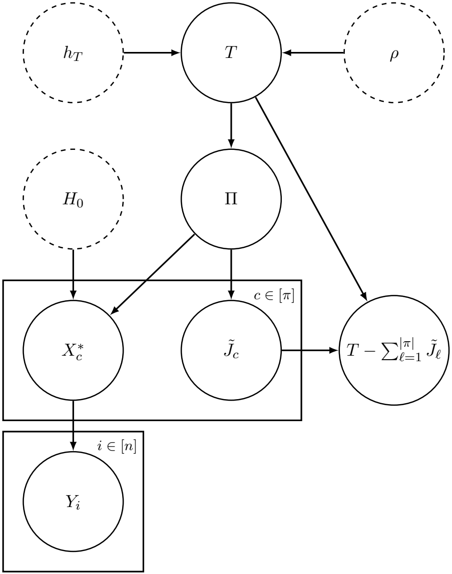


Figure: Tractable graphical model with additional auxiliary variables for an infinite mixture model

Auxiliary SMC sampler

[Lomeli, 2017] for Gibbs type priors, [Griffin, 2011] for Normalised Random Measure mixture models

Auxiliary SMC proposal and incremental weight

$$\Pr\left(\text{i joins cluster c'} \mid \Pi^{\ell}_{i-1}, \mathbf{y}_{1:i-1}, \left\{\tilde{J}_k \in \mathrm{d}s_k\right\}_{k=1}^{|\Pi^{\ell}_{i-\mathbf{1}}|}, T - \sum_{\ell \leq |\Pi^{\ell}_{i-\mathbf{1}}|} \tilde{J}_{\ell} \in \mathrm{d}v\right)$$

$$\propto \begin{Bmatrix} s_{c'} f\left(y_i\right) \mid \{y_j\}_{j \in c'}\right) & \text{if } c' \in \Pi^{\ell}_{i-1} \\ v f\left(y_i\right) & \text{o.w.} \end{Bmatrix}$$

and the incremental weight is

$$p\left(y_i \mid \Pi^{\ell}_{i}, \mathbf{y}_{1:i-1}, \left\{\tilde{J}_k \in \mathrm{d}s_k\right\}_{k=1}^{|\Pi^{\ell}_{i-\mathbf{1}}|}, T - \sum_{\ell \leq |\Pi^{\ell}_{i-\mathbf{1}}|} \tilde{J}_{\ell} \in \mathrm{d}v\right)$$

$$= \frac{v}{t} f(y^i) + \sum_{c \in \Pi^{\ell}_{i}} \frac{s_c}{t} f(y^i \mid \{y_j\}_{j \in c}).$$

SMC sampler: cluster assignment step

For the ℓ-th particle, in the PY process case,
$T^{\ell} \sim$ Polynomially tilted Stable(θ, S_σ), S_σ is a σ-Stable random variable.

[Devroye, 2009, Hofert, 2011]

SMC sampler: cluster assignment step

The ℓ-th particle (with no resampling step):
$\Pi_1^\ell = \{\{1\}\}$, $\mathbf{S}^\ell = \left[\tilde{J}_1\right]$, $V^\ell = T - \tilde{J}_1$.
[Lomeli, 2017]

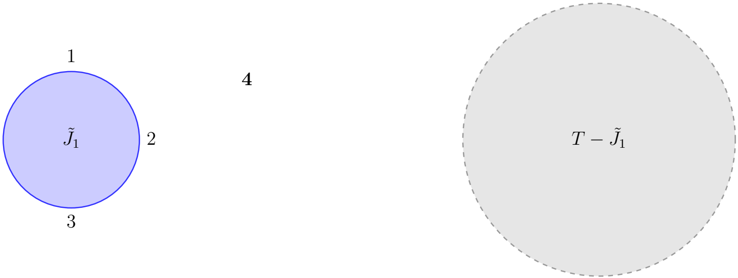
SMC sampler: cluster assignment step

The ℓ-th particle (with no resampling step):
$\Pi_2^\ell = \{\{1,2\}\}$, $\mathbf{S}^\ell = \left[\tilde{J}_1\right]$, $V^\ell = T - \tilde{J}_1$.
[Lomeli, 2017]

SMC sampler: cluster assignment step

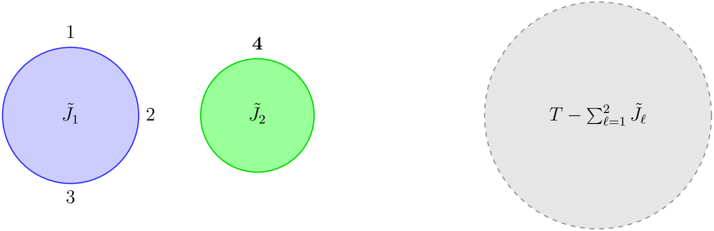
The ℓ-th particle (with no resampling step):
$\Pi_3^\ell = \{\{1,2,3\}\}$, $\mathbf{S}^\ell = \left[\tilde{J}_1\right]$, $V^\ell = T - \tilde{J}_1$.
[Lomeli, 2017]

SMC sampler: cluster assignment step



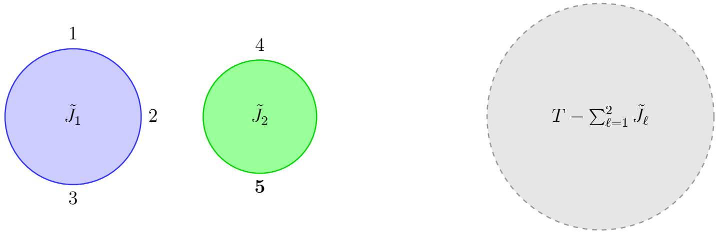
[Lomeli, 2017]

SMC sampler: cluster assignment step



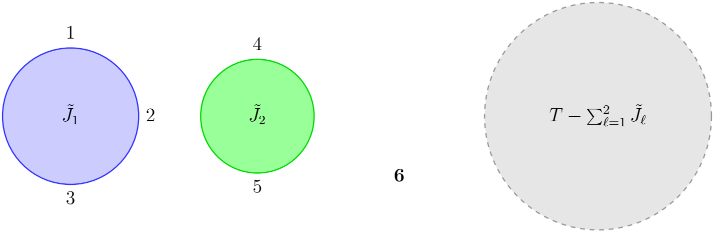
The ℓ-th particle (with no resampling step):
$\Pi_4^\ell = \{\{1,2,3\},\{4\}\}$, $\mathbf{S}^\ell = \left[\tilde{J}_1, \tilde{J}_2\right]$, $V^\ell = T - \tilde{J}_1 - \tilde{J}_2$.
[Lomeli, 2017]

SMC sampler: cluster assignment step



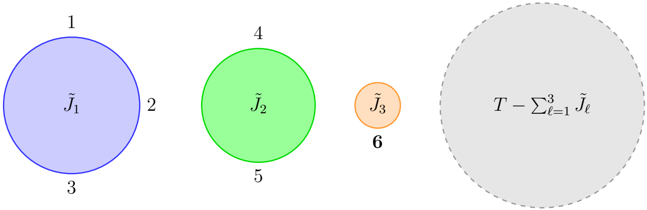
The ℓ-th particle (with no resampling step):
$\Pi_5^\ell = \{\{1,2,3\},\{4,5\}\}$, $\mathbf{S}^\ell = \left[\tilde{J}_1, \tilde{J}_2\right]$, $V^\ell = T - \tilde{J}_1 - \tilde{J}_2$.
[Lomeli, 2017]

SMC sampler: cluster assignment step



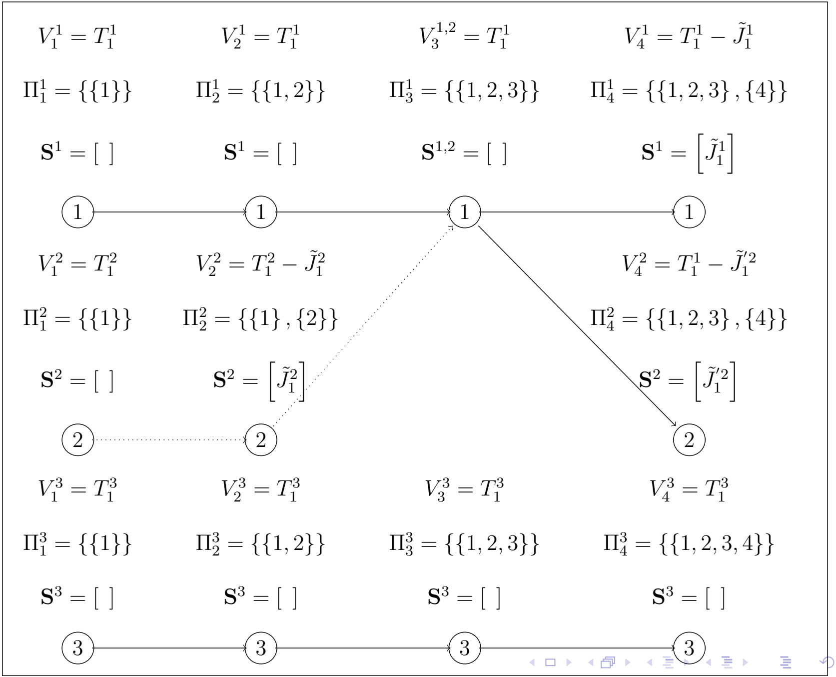
[Lomeli, 2017]

SMC sampler: cluster assignment step



The ℓ-th particle (with no resampling step):
$\Pi_6^\ell = \{\{1,2,3\},\{4,5\},\{6\}\}$, $\mathbf{S}^\ell = \left[\tilde{J}_1, \tilde{J}_2, \tilde{J}_3\right]$, $V^\ell = T - \tilde{J}_1 - \tilde{J}_2 - \tilde{J}_3$.
[Lomeli, 2017]

With a resampling step



Marginal likelihood computations

An advantage about using an SMC scheme is that the marginal likelihood can be directly estimated from the output by

$$\prod_{i=1}^{n} \frac{1}{L} \sum_{p=1}^{L} w_i^p.$$

This quantity is useful to construct a Bayes factor test.

Bayes factors

The Bayes factor allows us to compare the predictions made by two competing scientific theories represented by two statistical models.

$$\mathrm{BF} = \frac{p(\mathbf{D} \mid \mathcal{M}_1)}{p(\mathbf{D} \mid \mathcal{M}_2)}$$

where

$$p(\mathbf{D} \mid \mathcal{M}_k) = \int p(\mathbf{D} \mid \mathcal{M}_k, \phi_k) f(\phi_k \mid \mathcal{M}_k) d\phi_k, \quad k = 1, 2.$$

where $\mathbf{D} = (y_1, \cdots, y_n)$ is our data, $\mathcal{M}_1$ is model one, $\mathcal{M}_2$, model two; ϕ_k is the parameter under the hypothesis or competing model $\mathcal{M}_k$, $k = 1, 2$ and $f(\phi_k \mid \mathcal{M}_k)$ is its corresponding prior density.

[Jeffreys, 1935, Kass and Raftery, 1995, Robert, 2001]

Results

Bayes Factor

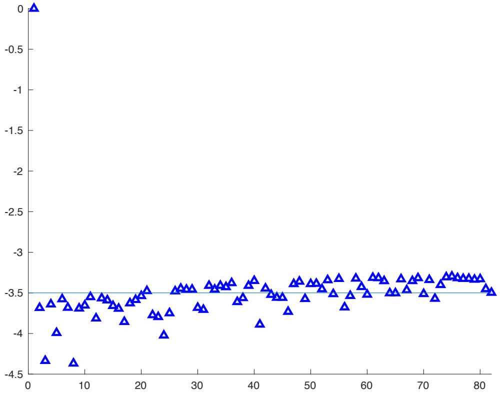


Figure: There is evidence in favour of the finite mixture model with random number of total components.

Conclusions

- From intractable to tractable representations useful for constructing inference schemes.
- SMC is a useful and general algorithm for inference in complex models.
- The SMC sampler presented is for a subclass of σ-Stable Poisson–Kingman mixture models. We have two other SMC samplers for Gibbs-type mixture models that were not covered here: one is an example of pseudo marginal MCMC and the other an approximate SMC scheme that encompasses all Gibbs type priors.

Main References

- Lomeli, M. **General Bayesian inference schemes in infinite mixture models.** PhD thesis, University College London, 2017.
- Favaro, S. and Lomeli, M. and Nipoti, B. and Teh, Y. W., **On the stick breaking representation of σ-Stable Poisson-Kingman models.** Electronic Journal of Statistics, 2014.

Thank you!

Aldous, D. (1985).
Exchangeability and related topics.
In *École d'Été de Probabilités de Saint-Flour XIII–1983*, pages 1–198. Springer, Berlin.

Beal, M. J., Ghahramani, Z., and Rasmussen, C. E. (2002).
The infinite hidden Markov model.
In *Advances in Neural Information Processing Systems*, volume 14.

Devroye, L. (2009).
Random variate generation for exponentially and polynomially tilted Stable distributions.
ACM Transactions on Modelling and Computer Simulation, 19:1–20.

Ghahramani, Z. (2015).
Probabilistic machine learning and artificial inteligence.
Nature, 521:452Ð459.

Griffin, J. (2011).
Sequential monte carlo methods for normalized random measures with independent increments mixtures.

Hofert, M. (2011).
Efficiently sampling nested archimedean copulas.
Comput. Statist. Data Anal., 55:57Ð70.

Jeffreys, H. (1935).
Some tests of significance treated by the theory of probability.
Mathematical Proceedings of the Cambridge Philosophical Society, 31:203–222.

Kass, R. E. and Raftery, E. (1995).
Bayes factors.
American Statistical Association, 90:773–795.

Kingman, J. F. C. (1975).
Random discrete distributions.
Journal of the Royal Statistical Society, 37:1–22.

Laha, R. G. and Rohatgi, V. (1979).
Probability theory.
John Wiley and Sons.

Lo, A. (1984).
On a class of bayesian nonparametric estimates: I. density estimates.
Annals of Statistics, 12:351–357.

Lomeli, M. (2017).
General Bayesian inference schemes in infinite mixture models.
PhD thesis, University College London.

Orbanz, P. (2014).
Lecture notes on bayesian nonparametrics.

Pitman, J. and Yor, M. (1997).
The two parameter Poisson-Dirichlet distribution derived from a Stable subordinator.
Annals of Probability, 25:855–900.

Rasmussen, C. E. (2000).
The infinite Gaussian mixture model.
In *Advances in Neural Information Processing Systems*, volume 12.

Robert, C. (2001).
The Bayesian Choice.
Springer.

Sethuraman, J. (1994).
A constructive definition of Dirichlet priors.
Statistica Sinica, 4:639–650.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2004).
Hierarchical Dirichlet processes.
Technical Report 653, Department of Statistics, University of California at Berkeley.

$\Pi_1^\ell = \{\{1\}\}, \forall \ell \in \{1, \cdots, L\}$

Sample $\mathbf{T}$=**GenerallyTiltedStable**(h_t, σ, L),

$\tilde{\mathbf{J}}^1$=**ExactSampleNewTableSize**(T, σ, L)

for $i = 2 : n$ **do**

  for $\ell = 1 : L$ **do**

    Set c' according to

$$\Pr\left(\text{i joins cluster c'} \mid \Pi_{i-1}^\ell, \mathbf{y}_{1:i-1}, \left\{\tilde{J}_k \in \mathrm{d}s_k\right\}_{k=1}^{|\Pi_{i-1}^\ell|}, T - \sum_{\ell \leq |\Pi_{i-1}^\ell|} \tilde{J}_\ell \in \mathrm{d}v\right)$$

    if $|c'| = 1$ **then**

      $\Pi_i^\ell = \Pi_{i-1}^\ell \cup \{\{i\}\}$

      $\tilde{J}_{|\Pi_i^\ell|}$=**ExactSampleNewTableSize**$\left(V := T - \sum_{\ell \leq |\Pi_{i-1}^\ell|} \tilde{J}_\ell, \sigma\right)$

      $V = V - \tilde{J}_{|\Pi_i^\ell|}$

    else

      $c' = c' \cup \{i\}, \quad c' \in \Pi_{i-1}^\ell$

      $\Pi_i^\ell = \Pi_{i-1}^\ell$

    end if

$$w_i^\ell \propto w_{i-1}^\ell \times p\left(y_i \mid \Pi_i^\ell, \mathbf{y}_{1:i-1}, \left\{\tilde{J}_k \in \mathrm{d}s_k\right\}_{k=1}^{|\Pi_{i-1}^\ell|}, T - \sum_{\ell \leq |\Pi_{i-1}^\ell|} \tilde{J}_\ell \in \mathrm{d}v\right)$$

  end for

  Normalise the weights $\tilde{w}_i^\ell = \frac{w_i^\ell}{\sum_{j=1}^L w_i^j}$

  if ESS$<$ thresh $\times L$ **then**

    Resample $\ell' \sim \text{Multinomial}\left(\tilde{w}_i^1, \cdots, \tilde{w}_i^L\right), \forall \ell \in \{1, \cdots, L\}, \ \Pi_i^\ell = \Pi_i^{\ell'}$

  end if

end for

Algorithm	Running time($\pm$std)	log-Marginal likelihood($\pm$std)
PY($\theta = 10, \sigma = 0.5$)		
StandardVanillaSMC	377.927 (35.29)	-294.622 (0.76)
StandardSMC	445.839 (15.65)	-292.704 (0.65)
VanillaSMC I	663.909 (39.36)	-297.865 (1.45)
SMC I	649.042 (32.03)	-298.129 (0.86)
ApproxVanillaSMC	543.429 (40.53)	-299.966 (0.50)
AproxSMC	420.818 (23.38)	-295.093 (0.47)
NGG($\tau = 20, \sigma = 0.5$)		
VanillaSMC I	417.735 (13.60)	-286.591 (0.14)
SMC I	429.590 (32.93)	-286.577 (0.35)
ApproxVanillaSMC	568.531 (29.29)	-299.149 (0.02)
ApproxSMC	511.341 (21.18)	-297.107 (0.13)
MFM($M \sim$ Gnedin($\gamma = 0.5$))		
VanillaSMC III	433.536 (135.82)	-276.129 (0.77)
SMC III	412.625 (116.99)	-276.427 (0.32)

Table: Running times in seconds and log-marginal likelihood averaged over 5 runs, 10000 particles.

Size-biased and Stick breaking weights for the Pitman–Yor process

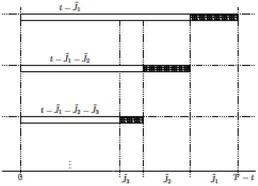


Figure 1: Generative process of Section 2.1

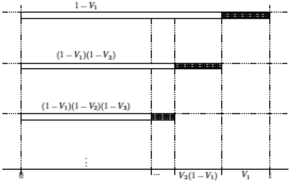


Figure 2: Pitman-Yor's stick breaking construction

$$
\begin{aligned}
T &\sim \gamma_{\mathrm{PY}} \\
\tilde{J}_1 \mid T &\sim \mathrm{SBS}(T) \\
\tilde{J}_2 \mid T, \tilde{J}_1 &\sim \mathrm{SBS}\left(T - \tilde{J}_1\right) \\
&\vdots \\
\tilde{J}_\ell \mid T, \tilde{J}_1, \ldots, \tilde{J}_{\ell-1} &\sim \mathrm{SBS}\left(T - \sum_{i<\ell} \tilde{J}_i\right) \\
&\vdots \\
P_\ell &\stackrel{d}{=} \frac{\tilde{J}_\ell}{T - \sum_{j<\ell} \tilde{J}_j}
\end{aligned}
$$

$$
\begin{aligned}
V_1 &\sim \mathrm{Beta}(v_1 \mid 1 - \sigma, \theta + \sigma) \\
V_2 &\sim \mathrm{Beta}(v_2 \mid 1 - \sigma, \theta + 2\sigma) \\
&\vdots \\
V_\ell &\sim \mathrm{Beta}(v_\ell \mid 1 - \sigma, \theta + \ell\sigma) \\
&\vdots
\end{aligned}
$$

the corresponding weights are:

$$P_\ell \stackrel{d}{=} V_\ell \prod_{j<\ell} (1 - V_j).$$