Gibbs Sampling Methods for Stick-Breaking Priors

Hermant Ishwaran
Lancelot F. James
JASA 2001

Presented by Daniel Johnson, Geoffrey Sun

Overview

- General class of *stick-breaking priors*
- Truncation result
- *Polya urn Gibbs sampler*
- *Blocked Gibbs sampler*
- Comparison

Stick-Breaking Priors

- Discrete random probability measures

$$\mathcal{P}(\cdot) = \sum_{k=1}^{N} p_k \delta_{Z_k}(\cdot), \qquad 0 \le p_k \le 1 \text{ and } \textstyle\sum_{k=1}^{N} p_k = 1 \text{ almost surely}$$

- Random measure $\mathbf{P}_N(\mathbf{a},\mathbf{b})$ is *stick-breaking random measure*

$$p_1 = V_1 \quad \text{and} \quad p_k = (1 - V_1)(1 - V_2) \cdots (1 - V_{k-1}) V_k, \quad k \ge 2,$$

- V_k ~ Beta(a_k,b_k), independent

- N finite <u>or</u> infinite

The Case N < Infinity

$$p_1 = V_1 \quad \text{and} \quad p_k = (1 - V_1)(1 - V_2) \cdots (1 - V_{k-1}) V_k, \quad k = 2, \ldots, N.$$

- N-1 degrees of freedom

Setting $V_N = 1$ guarantees that $\sum_{k=1}^{N} p_k = 1$ with probability 1, because

$$1 - \sum_{k=1}^{N-1} p_k = (1 - V_1) \cdots (1 - V_{N-1}). \tag{4}$$

The Case N = Infinity

$$p_1 = V_1 \quad \text{and} \quad p_k = (1 - V_1)(1 - V_2) \cdots (1 - V_{k-1}) V_k, \quad k \geq 2,$$

- Necessary and sufficient conditions

Lemma 1. For the random weights in the $\mathcal{P}_\infty(\mathbf{a}, \mathbf{b})$ random measure,

$$\sum_{k=1}^{\infty} p_k = 1 \quad \text{a.s. iff} \quad \sum_{k=1}^{\infty} E\big(\log(1 - V_k)\big) = -\infty. \qquad (5)$$

Alternatively, it is sufficient to check that $\sum_{k=1}^{\infty} \log(1 + a_k / b_k) = +\infty$.

- Computation?

The Pitman-Yor Process **PY**(a,b)

- Special case of *stick-breaking prior* $\mathbf{P}_N(\mathbf{a},\mathbf{b})$
- $a_k = 1 - a \qquad 0 <= a < 1$
- $b_k = b + ka \qquad b > -a$

Notable Pitman-Yor processes:

- Dirichlet process a = 0, b = alpha
- 'Stable law' process a = alpha, b = 0

Generalized Polya Urn Characterization

For a P-Y process **PY**(a,b)

$$\mathbb{P}\{Y_i \in \cdot \mid Y_1, \ldots, Y_{i-1}\} = \frac{b + am_i}{b + i - 1} H(\cdot) + \sum_{j=1}^{m_i} \frac{n^*_{j,i} - a}{b + i - 1} \delta_{Y^*_{j,i}}(\cdot), \qquad i = 2, 3, \ldots, n,$$

$\zeta_1, \ldots, \zeta_n$ iid H

$$Y_1 = \zeta_1,$$

$$(Y_i \mid Y_1, \ldots, Y_{i-1}) = \begin{cases} \zeta_i & \text{with probability } (b + am_i)/(b + i - 1), \\ Y^*_{j,i} & \text{with probability } (n^*_{j,i} - a)/(b + i - 1), \end{cases}$$

for $i = 2, 3, \ldots, n$.

Finite Dimensional Dirichlet **DP**$_N$(alpha*H)

- Like a Pitman-Yor process with
  - a = -alpha/N
  - b = alpha > 0
  - N >= n

$$\mathcal{P}(\cdot) = \sum_{k=1}^{N} \frac{G_{k,N}}{\sum_{k=1}^{N} G_{k,N}} \delta_{Z_k}(\cdot), \quad G_{k,N} \overset{\text{iid}}{\sim} \text{Gamma}\left(\frac{\alpha}{N}\right).$$

- **P**$_N$(**a,b**) measure, but not actually P-Y as (a < 0)

$$\text{DP}_N(\alpha H)(g) \xrightarrow{\text{d}} \text{DP}(\alpha H)(g)$$

Truncation of P_∞(**a,b**) Measures

- Truncations P_N(**a,b**) are computationally tractable
- Produce virtually indistinguishable measures

Theorem 2. Let p_k denote the random weights from a given $\mathcal{P}_\infty(\mathbf{a},\mathbf{b})$ measure. If $\|\cdot\|_1$ denotes the $\mathcal{L}_1$ distance, then

$$\|\mu_N - \mu_\infty\|_1 \le 4\left(1 - \mathrm{E}\left[\left(\sum_{k=1}^{N-1} p_k\right)^n\right]\right).$$

- Ex. Dirichlet Process

$$\|\mu_N - \mu_\infty\|_1 \sim 4n\exp(-(N-1)/\alpha)$$

Proof of Theorem 2

Proof sketch on board if time & interest.

Summary

- Stick-breaking prior
- Pitman-Yor
- Truncation

Two Gibbs Samplers

Polya Urn Gibbs Sampler

- extension to models from Escobar, West, MacEachern, Ferguson.
- integrates out P in the hierarchical model
- must have a known prediction rule ($Y_i|\mathbf{Y}_{-i}$)
- P can be infinite

Blocked Gibbs Sampler

- works when prediction rule is unknown
- directly involve the prior in the sampler
- *P* needs to be finite (but we can apply truncation for infinite *P*)

Polya Urn Gibbs Samplers

hierarchical model with stick-breaking priors:

$$
\begin{aligned}
(X_i|Y_i,\theta) &\stackrel{\text{ind}}{\sim} \pi(X_i|Y_i,\theta), \qquad i=1,\ldots,n,\\
(Y_i|P) &\stackrel{\text{iid}}{\sim} P,\\
\theta &\sim \pi(\theta),\\
P &\sim \mathcal{P}, \qquad (11)
\end{aligned}
$$

integrate out *P* :

CRP

$$
\begin{aligned}
(X_i|Y_i,\theta) &\stackrel{\text{ind}}{\sim} \pi(X_i|Y_i,\theta), \qquad i=1,\ldots,n,\\
(Y_1,\ldots,Y_n) &\sim \pi(Y_1,\ldots,Y_n),\\
\theta &\sim \pi(\theta), \qquad (12)
\end{aligned}
$$

Polya Urn Gibbs Samplers (PG)

- Assume priors = ***PY***(a,b) or ***DP***$_N$(alpha*H)
- Know the prediction rule ($Y_i|\mathbf{Y}_{-i}$)
- Want the posterior □(**Y**,theta|**X**)
- Iterate between the two steps:

  (a) ($Y_i|\mathbf{Y}_{-i}$,theta,**X**):

$$\mathbb{P}\{Y_i \in \cdot \mid \mathbf{Y}_{-i}, \theta, \mathbf{X}\}$$

assigned clusters

$$= q_0^* \, \mathbb{P}\{Y_i \in \cdot \mid \theta, X_i\} + \sum_{j=1}^{m} q_j^* \, \delta_{Y_j^*}(\cdot),$$

unseen clusters

$$q_0^* \propto (b + am) \int_y f(X_i|Y, \theta)\, H(dY),$$

$$q_j^* \propto (n_j^* - a)\, f(X_i|Y_j^*, \theta)$$

  (b) (theta|**Y**,**X**):

$$f(\theta|\mathbf{Y}, \mathbf{X}) \propto \pi(d\theta) \prod_{i=1}^{n} f(X_i|Y_i, \theta)$$

Polya Urn Gibbs Samplers (PG_a)

- Problem: **$\mathbf{Y}^*$** (unique Y's) get stuck if q_0^* is large
- Acceleration step: resample **$\mathbf{Y}^*$**
- Let **C**=$(C_1,\ldots,C_n)$, indexing into **$\mathbf{Y}^*$** as a look-up table

(c) $(Y_j^*|\mathbf{C},\text{theta},\mathbf{X})$

$$f(Y_j^*|\mathbf{C},\theta,\mathbf{X}) \propto H(dY_j^*) \prod_{\{i: C_i = j\}} f(X_i|Y_j^*,\theta).$$

Limitations of PG and PG_a

1. slow mixing: a single Y_i at a time
2. relies on conjugacy for q_0^*
3. prior *P* is not directly involved, only depend on **Y**
4. *requires a known urn scheme / prediction rule*

Blocked Gibbs Sampler

- need finite prior $P_N(a,b)$ (use trucations as in section 3.2)
- update blocks of parameters (draw from multivariate distr)

$$
\begin{aligned}
(X_i|\mathbf{Z},\mathbf{K},\theta) &\overset{\text{ind}}{\sim} \pi(X_i|Z_{K_i},\theta), \quad i=1,\ldots,n, \\
(K_i|\mathbf{p}) &\overset{\text{iid}}{\sim} \sum_{k=1}^{N} p_k\,\delta_k(\cdot), \\
(\mathbf{p},\mathbf{Z}) &\sim \pi(\mathbf{p})\times H^N(\mathbf{Z}), \\
\theta &\sim \pi(\theta),
\end{aligned}
$$

$Y_i = Z_{Ki}$

where $\mathbf{K} = (K_1,\ldots,K_n)$, $\mathbf{Z} = (Z_1,\ldots,Z_N)$, $\mathbf{p} = (p_1,\ldots,p_N) \sim \mathcal{GD}(\mathbf{a},\mathbf{b})$, and Z_k are iid H.

Blocked Gibbs Sampler

direct posterior inference:

iterate

$$(\mathbf{Z}|\mathbf{K}, \theta, \mathbf{X}),$$

$$(\mathbf{K}|\mathbf{Z}, \mathbf{p}, \theta, \mathbf{X})$$

$$(\mathbf{p}|\mathbf{K}),$$

$$(\theta|\mathbf{Z}, \mathbf{K}, \mathbf{X}).$$

equilibrium distribution

$$(\mathbf{Z}, \mathbf{K}, \mathbf{p}, \theta|\mathbf{X})$$

each draw gives a random measure

$$P(\cdot) = \sum_{k=1}^{N} p_k \, \delta_{Z_k}(\cdot),$$

Blocked Gibbs Sampler

(a) Conditional for $\mathbf{Z}$: Simulate $Z_k \overset{\text{iid}}{\sim} H$ for each $k \in \mathbf{K} - \{K_1^*, \ldots, K_m^*\}$. Also, draw $(Z_{K_j^*}|\mathbf{K}, \theta, \mathbf{X})$ from the density

$$f(Z_{K_j^*}|\mathbf{K}, \theta, \mathbf{X}) \propto H(dZ_{K_j^*}) \prod_{\{i: K_i = K_j^*\}} f(X_i|Z_{K_j^*}, \theta), \quad j = 1, \ldots, m. \quad (18)$$

(b) Conditional for $\mathbf{K}$: Draw values

$$(K_i|\mathbf{Z}, \mathbf{p}, \theta, \mathbf{X}) \overset{\text{ind}}{\sim} \sum_{k=1}^{N} p_{k,i}\, \delta_k(\cdot), \qquad i = 1, \ldots, n,$$

where

$$(p_{1,i}, \ldots, p_{N,i}) \propto \big(p_1 f(X_i|Z_1, \theta), \ldots, p_N f(X_i|Z_N, \theta)\big).$$

(a) ~ acceleration step (c)

(d) = (b) in PG and PG_a

(c) Conditional for $\mathbf{p}$: By the conjugacy of the generalized Dirichlet distribution to multinomial sampling, it follows that our draw is

$$p_1 = V_1^* \quad \text{and} \quad p_k = (1 - V_1^*)(1 - V_2^*) \cdots (1 - V_{k-1}^*) V_k^*, \quad k = 2, \ldots, N-1,$$

where

$$V_k^* \overset{\text{ind}}{\sim} \text{Beta}\Big(a_k + M_k, b_k + \sum_{l=k+1}^{N} M_l\Big), \quad \text{for } k = 1, \ldots, N-1,$$

(d) Conditional for θ: As before, draw θ from the density (remembering that $Y_i = Z_{K_i}$)

$$f(\theta|\mathbf{Z}, \mathbf{K}, \mathbf{X}) \propto \pi(d\theta) \prod_{i=1}^{n} f(X_i|Y_i, \theta).$$

Evaluation

- {DP, DP_{50}, PY} x {PG, PG_a, BG}
- experimental results (batch means, std, etc)

- PG is bad
- PG_a works well when prediction rule is known
- BG is more flexible

- complexity? linear in *n* (PG) vs multivariate draws (BG)
- easy to get stuck?
- which sampler for which prior?

Discussions

- PG : PG_a is the same as Algorithm 1 : 2 in [Neal 1999]

- what's semiparametric?

- what's "almost sure"?

- what's the graphical model like for (16) in Section 5?

- in the non-conjugate case, are we doing Metropolis-Hastings inside the Gibbs sampler? (Section 5.4)

- with known prediction rule, is BG preferred? How?

- sampling in equivalence class space vs label space?