Exchangeable random measures and stick-breaking priors

Ramsés H. Mena
O'Bayes, 2022

IIMAS-UNAM, México

(work with María F. Gil-Leyva, Theodoros Nicoleris, Antonio Lijoi and Igor Prünster)

September 9, 2022

Talk plan

1. Stick-breaking priors
2. Exchangeble SB prior
3. Markov stick-breaking processes
4. ESB-Mixture model

The basic setup

Random phenomena encoded in $\{X_i\}_{i=1}^{\infty}$ r.v.'s

- Statistical learning requires stochastic dependence !
  - ▷ Logical/physical independence $\not\Rightarrow$ stochastic independence

    so $\mathbb{P}(X_{n+1} \in B \mid X_1, \ldots, X_n) = \mathbb{P}(X_{n+1} \in B)$ not always a good idea!
  - ▷ Under physical independence of obs. all we can assume is certain stochastic symmetry among $\{X_i\}$

The basic setup

Random phenomena encoded in $\{X_i\}_{i=1}^{\infty}$ r.v.'s

- Statistical learning requires stochastic dependence !
  - ▷ Logical/physical independence $\not\Rightarrow$ stochastic independence

    so $\mathbb{P}(X_{n+1} \in B \mid X_1, \ldots, X_n) = \mathbb{P}(X_{n+1} \in B)$ not always a good idea!

  - ▷ Under physical independence of obs. all we can assume is certain stochastic symmetry among $\{X_i\}$
- **Exchangeability**

$$(X_1, \ldots, X_n) \stackrel{\mathrm{d}}{=} (X_{\pi(1)}, \ldots, X_{\pi(n)}), \qquad \forall n \geq 1$$

and for any permutation π of $\{1, \ldots, n\}$.

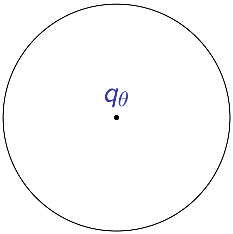
$\approx$ Distributional invariance under sampling order

Exchangeability

$\mathbb{X}$-valued $\{X_i\}_{i=1}^{\infty}$ exchangeable sequence driven by $\mathrm{P} \sim \mathrm{Q}$

- $\mathrm{Q}(\cdot) = \delta_{q_\theta}(\cdot) \Rightarrow X_i$'s are iid

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \int_{\mathcal{P}_\mathbb{X}} \prod_{i=1}^{n} \mathrm{P}(A_i)\, \delta_{q_\theta}(d\mathrm{P}) = \prod_{i=1}^{n} q_\theta(A_i)$$



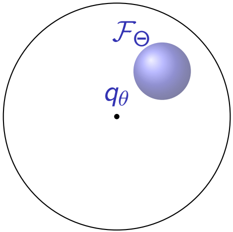
$\mathcal{P}_\mathbb{X}$: Space of all distributions on $\mathbb{X}$

Exchangeability

$\mathbb{X}$-valued $\{X_i\}_{i=1}^{\infty}$ exchangeable sequence driven by $\mathrm{P} \sim \mathrm{Q}$

- $\mathrm{Q}(\mathcal{F}_\Theta) = 1 \Rightarrow$ Parametric family

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \int_{\mathcal{F}_\Theta} \prod_{i=1}^{n} \underbrace{F_\theta(A_i)}_{\text{Random uncertainty via param. model}} \overbrace{\pi_\theta(d\theta)}^{\text{Epistemic uncertainty}}$$



$\mathcal{P}_{\mathbb{X}}$: Space of all distributions on $\mathbb{X}$

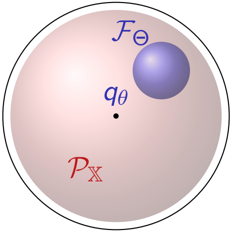
Exchangeability

$\mathbb{X}$-valued $\{X_i\}_{i=1}^{\infty}$ exchangeable sequence driven by $\mathrm{P} \sim \mathrm{Q}$

- $\mathrm{Q}(\mathrm{P} : \mathrm{d}(\mathrm{P}, \eta) < \varepsilon) > 0, \ \forall \ \eta \in \mathcal{P}_{\mathbb{X}}$ y $\varepsilon > 0 \Rightarrow$ BNP

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \int_{\mathcal{P}_{\mathbb{X}}} \prod_{i=1}^{n} \underbrace{\mathrm{P}(A_i)\, \mathrm{Q}(d\mathrm{P})}$$

Random and epistemic uncertainties in one stroke!

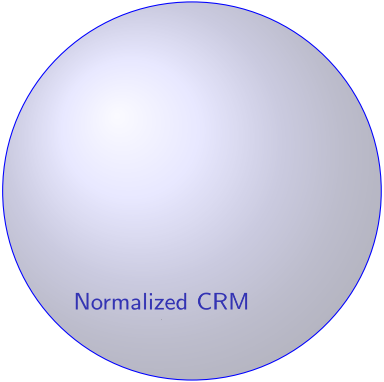


$\mathcal{P}_{\mathbb{X}}$: Space of all distributions on $\mathbb{X}$
...or other infinite dimensional sub-spaces of interest, $\mathcal{P}_{\mathbb{X}}^{d}$, $\mathcal{P}_{\mathbb{X}}^{c}$, etc.

How to construct suitable models for Q (nonparametric priors!)?

Popular constructions of discrete BNP priors

Given a CRM $\boldsymbol{\mu}$ satisfying $0 < \boldsymbol{\mu}(\mathbb{X}) < \infty \Rightarrow \mathrm{P}(\cdot) = \frac{\boldsymbol{\mu}(\cdot)}{\boldsymbol{\mu}(\mathbb{X})}$

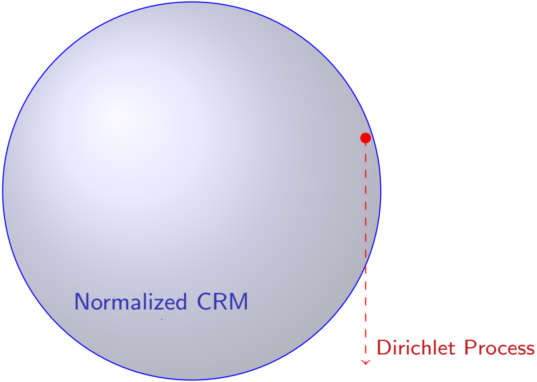


Prünster (2002), Regazzini *et. al.*(2003), James *et. al.* (2006)

Popular constructions of discrete BNP priors

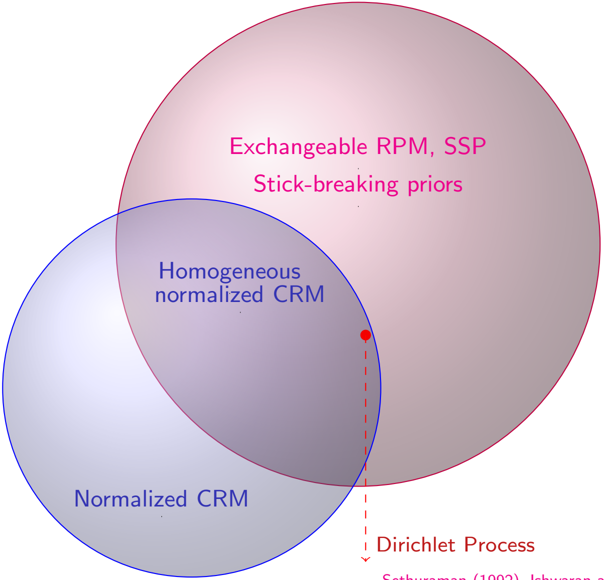
Given a CRM $\boldsymbol{\mu}$ satisfying $0 < \boldsymbol{\mu}(\mathbb{X}) < \infty \Rightarrow \mathrm{P}(\cdot) = \frac{\boldsymbol{\mu}(\cdot)}{\boldsymbol{\mu}(\mathbb{X})}$

If $\mathbb{E}[\boldsymbol{\mu}] = \nu(ds, dx) = s^{-1}e^{-s}\theta P_0(dx) \quad \Rightarrow \quad \mathrm{P} \sim \mathcal{D}(\theta P_0)$



Ferguson (1973)

Popular constructions of discrete BNP priors



$$P = \sum_{i\geq 1} w_i \delta_{\xi_i}$$

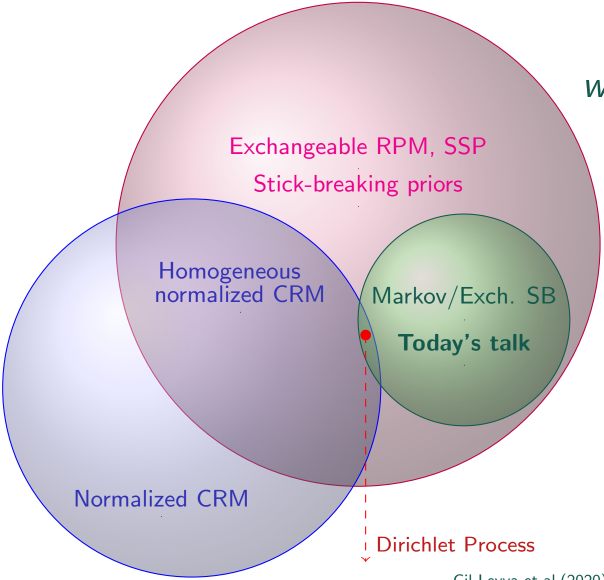
$$\sum_{i\geq 1} w_i = 1$$

$$\xi_i \overset{\text{iid}}{\sim} P_0$$

$$(w_i)_{i\geq 1} \perp (\xi_i)_{i\geq 1}$$

Sethuraman (1992), Ishwaran and James (2001), Kallenberg (2017)

Popular constructions of discrete BNP priors



$$w_i = v_i \prod_{j=1}^{i-1}(1 - v_j)$$

Dependent

Length variables

Gil-Leyva et al.(2020), Gil-Leyva and M. (2021)

Stick breaking weights

$$\mathrm{P}(B) = \sum_{i=1}^{\infty} w_i \, \delta_{\xi_i}(B), \quad B \in \mathcal{X}, \qquad \sum_i w_i = 1$$

with $w_i = v_i \prod_{j=1}^{i-1}(1 - v_j)$ and $\xi_i \overset{\text{iid}}{\sim} P_0$

- **Full support if:** For every $\varepsilon > 0$ there exist $m \in \mathbb{N}$ such that

$$\mathbb{P}[v_1 < \varepsilon, ..., v_m < \varepsilon] > 0$$

Stick breaking weights

$$\mathrm{P}(B) = \sum_{i=1}^{\infty} w_i \, \delta_{\xi_i}(B), \quad B \in \mathcal{X}, \qquad \sum_i w_i = 1$$

with $w_i = v_i \prod_{j=1}^{i-1}(1 - v_j)$ and $\xi_i \overset{\text{iid}}{\sim} P_0$

- **Full support if:** For every $\varepsilon > 0$ there exist $m \in \mathbb{N}$ such that

$$\mathbb{P}[v_1 < \varepsilon, \ldots, v_m < \varepsilon] > 0$$

- **Weights add up to one if**

$$\sum_{j \geq 1} w_j = 1 \quad \Leftrightarrow \quad \prod_{i=1}^{j}(1 - v_i) \overset{a.s}{\to} 0 \quad \Leftrightarrow \quad \mathbb{E}\left[\prod_{i=1}^{j}(1 - v_i)\right] \to 0$$

Some SB representations of well-known RPMs

- Dirichlet process: $v_i \stackrel{\text{iid}}{\sim} \text{Be}(1, \beta)$
- Two parameter Poisson-Dirichlet: $v_i \stackrel{\text{ind}}{\sim} \text{Be}(1 - \sigma, \beta + i\sigma)$
- σ-stable Poisson-Kingman: dependent $(v_i)_{i \geq 1}$ with

$$g(v_j \mid t, v_1, \ldots, v_{j-1}) = \frac{\sigma (tz_j)^{-\sigma}}{\Gamma(1-\sigma) f_\sigma(tz_j)} v_j^{-\sigma} f_\sigma(tz_j(1 - v_j))$$

  where f_σ denotes the positive σ stable density function and $z_j := \prod_{i=1}^{j-1}(1 - v_j)$ with $z_1 = 1$.

$\vdots$

- Homogeneous NRMIs... also dependent $(v_i)_{i \geq 1}$ with more involved conditional distributions for v_1 and $v_i \mid v_{i-1}, \ldots, v_1$

Favaro *et. al.* (2014), Favaro *et. al.* (2016).

BNP clustering structure

- Relies on "analytical" expressions of the EPPF, i.e. π s.t.

$$\mathbb{P}\left(\Pi(\mathbf{x_{1:n}}) = A\right) = \pi(n_1, \dots, n_k) = \sum_{(j_1,\dots,j_k)} \mathbb{E}\left[\prod_{i=1}^{k} w_{j_i}^{n_i}\right]$$

with $A = \{A_1, \dots, A_k\}$ a partition of $\mathbf{x_{1:n}}$ and $n_j := |A_j|$

$\Rightarrow$ Similar inference can be achieved via allocation variables, i.e.,
Given $\{x_i\}_{i\geq 1}$ exch. driven by a SSP $\mu = \sum w_j \delta_{\xi_j}$, $d_i = j$ iff $x_i = \xi_j$

$$\mathbb{P}\left(\mathbf{d_1} = d_1, \dots, \mathbf{d_n} = d_n\right) = \mathbb{E}\left[\prod_{j=1}^{k} v_j^{r_j}(1 - v_j)^{t_j}\right]$$

with $k := \max\{d_1, \dots, d_n\}$, $r_j := \sum_{i=1}^{n} \mathbf{1}_{(d_i = j)}$ and $t_j := \sum_{i=1}^{n} \mathbf{1}_{(d_i > j)}$.

Dirchlet process & Geometric process

Independent / Fully dependent random lengths

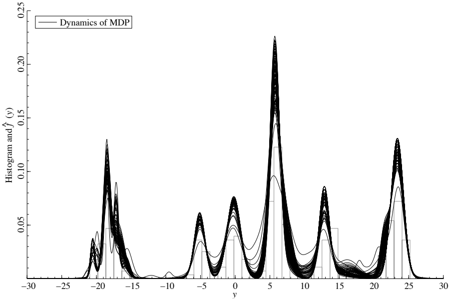
Dirichlet process	Geometric process
• $v_i \overset{\text{iid}}{\sim} \text{Be}(1, \beta)$	• $v_i = \lambda \sim \text{Be}(\alpha, \beta)$
• $w_j = v_j \prod_{i=1}^{j-1}(1 - v_i)$	• $w_j = \lambda(1 - \lambda)^{j-1}$
• $\mathbb{E}[w_1] > \mathbb{E}[w_2] > \cdots$	• $w_1 > w_2 > \cdots$
⋆ Size-biased random order	⋆ Decreasing order

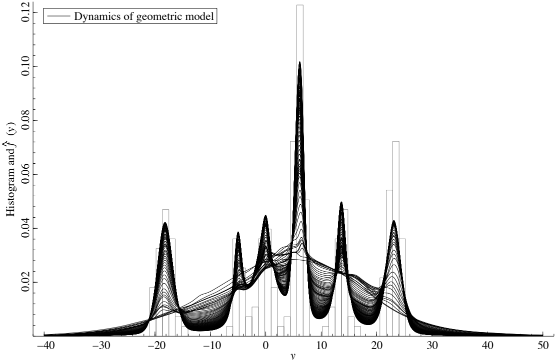
Both processes satisfy:

- $\sum_{j\geq 1} w_j = 1$.
- Have full support.
- Both are exchangeable!

⋆ We consider the general class of priors induced by exchangeable length variables

Ferguson (1973), Fuentes-García, M. and Walker (2010)


Dynamics of MDP
Histogram and $\hat{f}$ (y)
y
0.25
0.20
0.15
0.10
0.05
–30
–25
–20
–15
–10
–5
0
5
10
15
20
25
30


Dynamics of geometric model
Histogram and $\hat{f}$ (y)
y
0.12
0.10
0.08
0.06
0.04
0.02
–40
–30
–20
–10
0
10
20
30
40
50

Exchangeable stick-breaking processes

- ESB(ν, μ_0)

$$\mu = \sum_{j\geq 1} w_j \delta_{\xi_j} \xrightarrow{\text{atoms}} \xi_j \overset{\text{iid}}{\sim} \mu_0$$

with

$$w_j = v_j \prod_{i<j} (1 - v_i) \xrightarrow{\text{length var.}} (v_j) \text{ exchangeable seq. driven by } \nu$$

That is $v_j | \nu \overset{\text{iid}}{\sim} \nu_0$ with $\nu_0 := \mathbb{E}[\nu]$. (v_j) are $[0,1]$-valued.

Theorem

- $\nu(\{0\}) < 1$ a.s iff $\sum_{j\geq 1} w_j = 1$ a.s.
  - If $\nu_0(\{0\}) = 0$ then $\sum_{j\geq 1} w_j = 1$ a.s.
- If there exists $\epsilon > 0$ such that $(0, \epsilon)$ is contained in the support of ν_0, then μ has full support.

Convergence to Dirichlet and Geometric processes

$$\mathrm{ESB}(\nu, \mu_0)$$

$$\mathrm{DP}(\theta, \mu_0) \xleftarrow{\quad\mathrm{d}\quad} \mu \xrightarrow{\quad\mathrm{d}\quad} \mathrm{GP}(\nu_0, \mu_0)$$

$$\mathrm{Be}(1, \theta) \xleftarrow{\mathrm{d}} \nu \qquad\qquad \nu \xrightarrow{\mathrm{d}} \delta_v, \ \text{ with } v \sim_{\nu_0}$$

If $\nu = \sum_{j\geq 1} p_j \delta_{u_j}$ is a SSP with $\rho := \mathbb{P}[v_1 = v_2] = \sum_{j\geq 1} \mathbb{E}[p_j^2]$

$$\nu_0 \xleftarrow{\quad\mathrm{d}\quad} \nu \xrightarrow{\quad\mathrm{d}\quad} \delta_v, \ \text{ with } v \sim_{\nu_0}$$

$$0 \xleftarrow{\mathrm{d}} \rho \qquad\qquad \rho \xrightarrow{\mathrm{d}} 1$$

Take $\nu_0 = \mathrm{Be}(1, \theta)$. If $\nu \sim \mathrm{DP}(\beta, \nu_0)$, $\rho = \frac{1}{1+\beta}$.

Convergence to Dirichlet and Geometric processes

Ordering of the weights

$$\mu = \sum_{j\geq 1} w_j \delta_{\xi_j} \overset{d}{=} \sum_{j\geq 1} w_{\rho(j)} \delta_{\xi_j}$$

- One usually work with the ordering of weights that is the most tractable.

Size-biased DP weights

$$\mathbb{E}[\tilde{w}_1] > \mathbb{E}[\tilde{w}_2] > \cdots$$

Decreasing GP weights

$$w_1^{\downarrow} > w_2^{\downarrow} > \cdots$$

$$(\tilde{w}_j) \xleftarrow{\quad\mathrm{d}\quad} (w_j) \xrightarrow{\quad\mathrm{d}\quad} (w_j^{\downarrow})$$

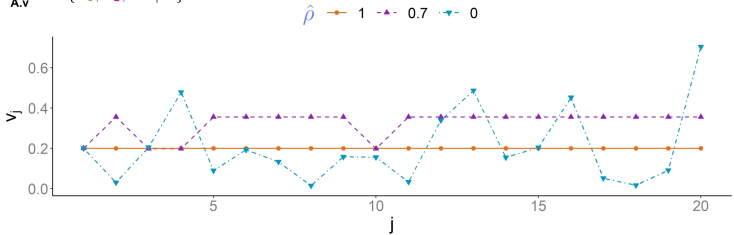
$$\mathrm{Be}(1,\theta) \xleftarrow{\mathrm{d}} \nu \qquad \nu \xrightarrow{\mathrm{d}} \delta_v, \text{ with } v \sim_{\nu_0}$$

Stick-breaking processes driven by SSP

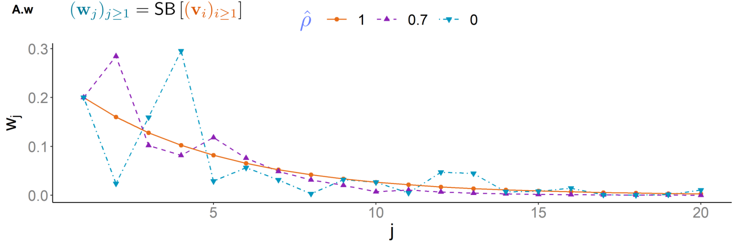
SSP — Base measure — $\nu_0 = \mathrm{Be}(1, \theta)$ — $\mathbf{v}_i \sim \nu_0$

SSP — Tie probability — $\hat{\rho} = \mathbb{P}[\mathbf{v}_1 = \mathbf{v}_2]$

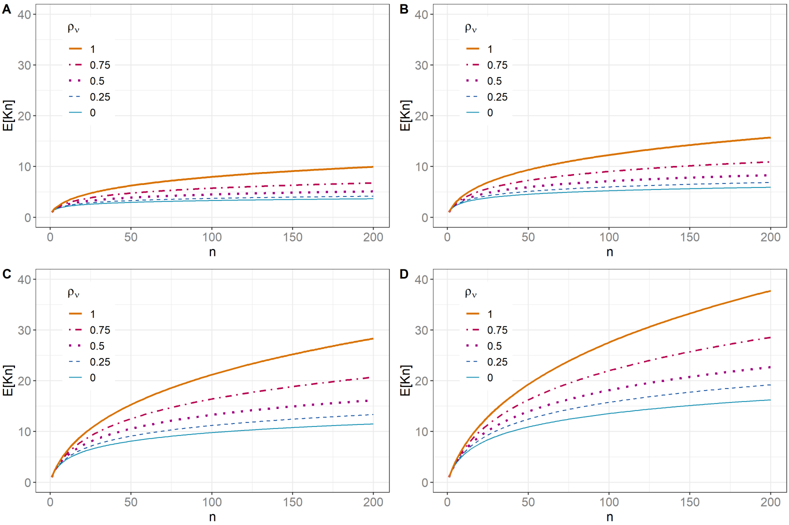
A.v $\{\mathbf{v}_1, \mathbf{v}_2, \ldots \mid \boldsymbol{\nu}\} \overset{iid}{\sim} \boldsymbol{\nu}$



A.w $(\mathbf{w}_j)_{j \geq 1} = \mathrm{SB}\left[(\mathbf{v}_i)_{i \geq 1}\right]$



Asymptotic behavior of $E[K_n]$



$\{A, B, C, D\}$ corresponds to $\theta = \{0.5, 1, 2.5, 4\}$

Plan

Exchangeable vs Markov stick-breaking process

$$\mu = \sum_{j\geq 1} w_j \delta_{\xi_j} \dashrightarrow \xi_j \overset{iid}{\sim} \mu_0$$

$$w_j = v_j \prod_{i<j}(1 - v_i)$$

Exchangeable

$$v_1 \quad v_2 \quad v_3 \quad \dots \quad v_j$$

$$\nu$$

ν takes values in $\mathcal{P}([0,1])$

$$\mathbb{E}[\nu] = \nu_0$$

$$v_j \sim \nu_0$$

$$\mu \sim \mathsf{ESB}(\nu, \mu_0)$$

Markov (stationary)

$$v_1 \xrightarrow[\nu]{} v_2 \xrightarrow[\nu]{} v_3 \xrightarrow[\nu]{} \cdots \xrightarrow[\nu]{} v_j$$

For each $u \in [0,1]$, $\nu(u, \cdot) \in \mathcal{P}([0,1])$

$$\mathbb{P}[v_{j+1} \in \cdot \mid v_j] = \nu(v_j, \cdot)$$

$$\nu_0(B) = \int \nu(u; B)\nu_0(du), \quad B \in \mathcal{B}([0,1])$$

$$v_1 \sim \nu_0 \quad \Rightarrow \quad v_j \sim \nu_0$$

$$\mu \sim \mathsf{MSB}(\nu, \mu_0)$$

Exchangeable vs. Markov stick-breaking process

Theorem (Exchangeable)

- $\nu(\{0\}) < 1$ a.s iff $\sum_{j\geq 1} w_j = 1$ a.s.
    - $*$ If $\nu_0(\{0\}) = 0$ then $\sum_{j\geq 1} w_j = 1$ a.s.
- If there exists $\epsilon > 0$ such that $(0, \epsilon)$ is contained in the support of ν_0, then μ has full support.

Theorem (Stationary Markov)

- $\nu_0 \neq \delta_0$ iff $\sum_{j\geq 1} w_j = 1$ a.s.
- If there exists $\epsilon > 0$ such that $(0, \epsilon)$ is contained in the support of ν_0, and for each $u \in (0, \epsilon)$, $(0, \epsilon)$ is contained in the support of $\nu(u, \cdot)$, then μ has full support.

Convergence to Dirichlet and Geometric processes

Size-biased DP weights

$\mathbb{E}[\tilde{w}_1] > \mathbb{E}[\tilde{w}_2] > \cdots$

Decreasing GP weights

$w_1^{\downarrow} > w_2^{\downarrow} > \cdots$

$$v_j \prod_{i<j} (1 - v_i)$$

$v_j \overset{iid}{\sim} \mathsf{Be}(1,\theta)$ $\qquad$ $v_j \mid \nu \overset{iid}{\sim} \nu$ $\qquad$ $v_j = v \sim \nu_0$

$$(\tilde{w}_j) \xleftarrow[\mathsf{Be}(1,\theta) \leftsquigarrow \nu]{d} (w_j) \xrightarrow[\nu \rightsquigarrow \delta_{\bullet}]{d} (w_j^{\downarrow})$$

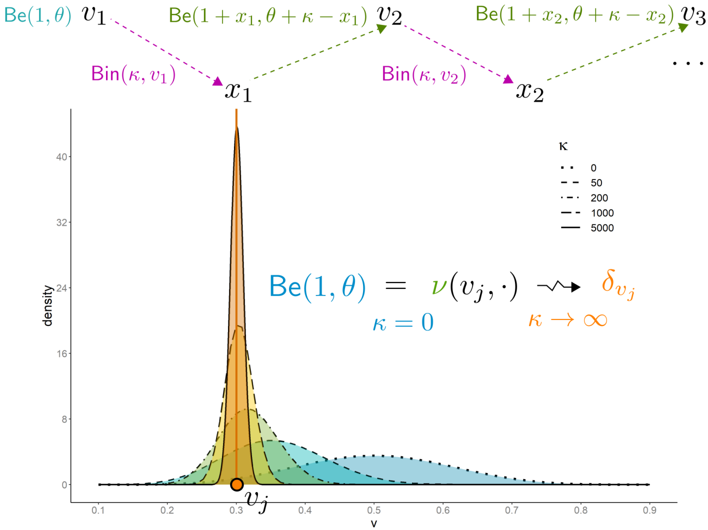
$\mathsf{MSB}(\nu, \mu_0)$

$$\mathsf{DP}(\theta, \mu_0) \xleftarrow[\mathsf{Be}(1,\theta) \leftsquigarrow \nu]{d} \mu \xrightarrow[\nu \rightsquigarrow \delta_{\bullet}]{d} \mathsf{GP}(\nu_0, \mu_0)$$

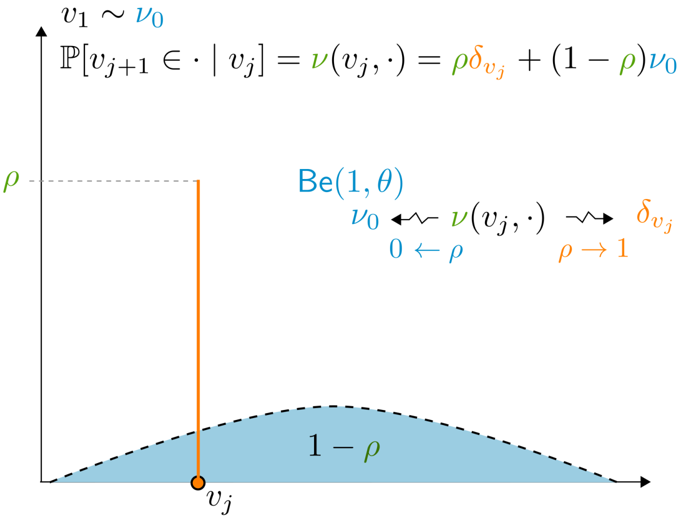
$$\boxed{\nu_n \rightsquigarrow \varphi \iff \begin{array}{c} \nu_n(u_n, \cdot) \overset{w}{\to} \varphi(u, \cdot) \\ \forall u_n \to u \text{ in } [0,1] \end{array}}$$

Beta-Binomial transition

$$\nu(v_j, \cdot) = \sum_{x=0}^{\kappa} \mathrm{Be}(\cdot \mid 1 + x, \theta + \kappa - x)\mathrm{Bin}(x \mid \kappa, v_j)$$



Spike and slab transition



Decreasing probability

If ν is a spike and slab transition

$$\mathbb{P}[w_{j+1} \leq w_j] = \rho + (1-\rho)\mathbb{E}\left[\overrightarrow{\nu_0}(c(v))\right]$$

If $\nu_0 = \mathrm{Be}(1, \theta)$

$$\mathbb{P}[w_{j+1} \leq w_j] = 1 - \frac{{}_2F_1(1, 1; \theta+2, 1/2)(1-\rho)\theta}{2(\theta+1)}$$

If $\theta = 1$

$$\mathbb{P}[w_{j+1} \leq w_j] = \rho + (1-\rho)\log(2)$$

Plan

ESB-Mixture model

$$\mu \sim \mathrm{ESB}(\nu, \mu_0)$$

- Data modeled via $y_i \mid \tilde{f} \overset{\text{iid}}{\sim} \tilde{f}$ for $i = 1, 2, \ldots$ with $\tilde{f}$ a μ-mixture, e.g.

$$\tilde{f}(y) = \int \mathrm{N}(y \mid x)\mu(dx) = \sum_{j \geq 1} w_j \, \mathrm{N}(y \mid \xi_j)$$

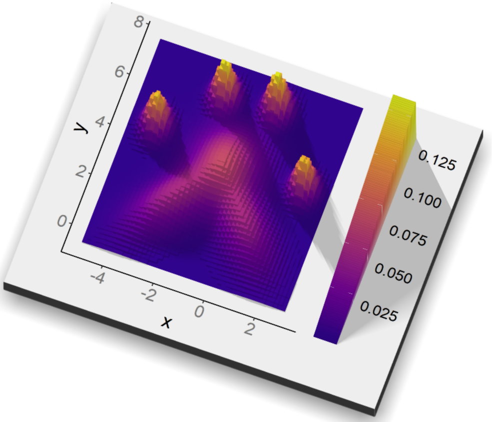
▷ $\mu_0 = \text{Normal} - \text{Inverse Wishart}$

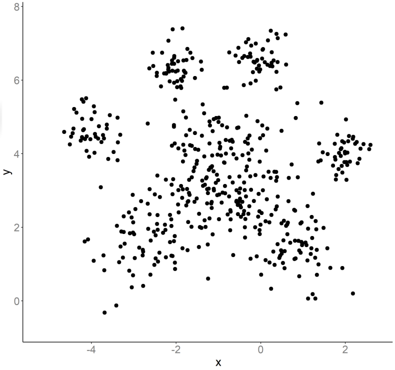
▷ $\nu_0 = \mathrm{Be}(1, \theta)$, $\theta = 1$

▷ $\rho \sim \mathrm{U}[0, 1]$

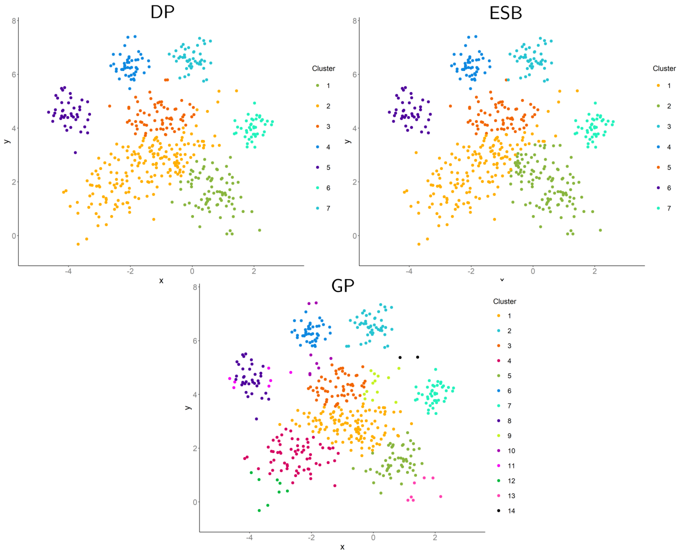
Gibbs sampler, e.g. Walker (2007), Kalli *et al.* (2011)

Paw dataset

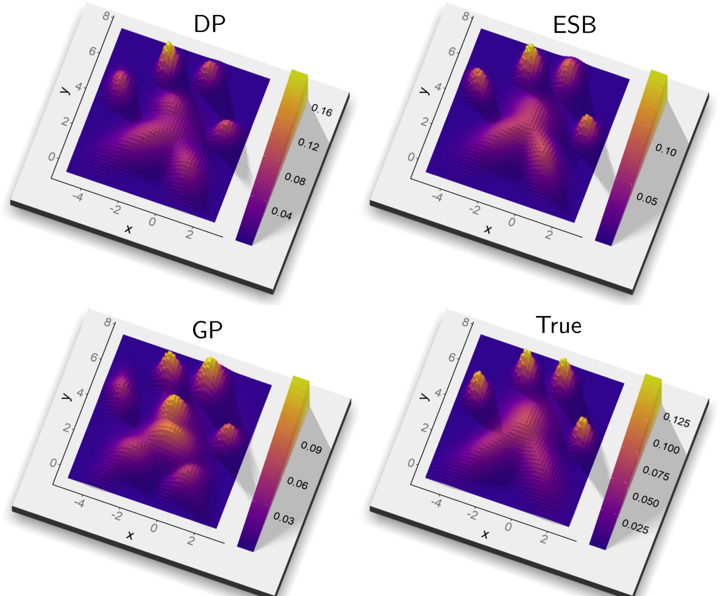




Cluster estimation for Paw dataset



Density estimation for Paw dataset



References

Favaro, S., Lijoi, A., Nava, C., Nipoti, B., Prünster, I. and Teh, Y. (2016). On the stick-breaking representation for homogeneous NRMIs, *Bayesian Analysis* 11: 697–724.

Favaro, S., Lomeli, M., Nipoti, B. and Teh, Y. (2014). On the stick-breaking representation of ?σ-stable Poisson-Kingman models, *Electronic Journal of Statistics* 8: 1063?1085.

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

Fuentes-García, R., Mena, R. H. and Walker, S. G. (2010). A new Bayesian nonparametric mixture model. *Communications in Statistics-Simulation and Computation*. 39, 669–682.

Gil-Leyva, M.F., Mena, R.H. and Nicoleris, T. (2020). Beta-Binomial stick-breaking non-parametric prior. *Electronic Journal of Statistics*. 14, 1479–1507.

Gil-Leyva, M.F. and Mena, R.H. (2021). Stick-breaking processes with exchangeable length variables. *Journal of the American Statistical Association*. To appear.

Mena, R. and Walker, S. G. (2009). On a construction of Markov models in continuous time, *METRON–International Journal of Statistics* LXVII: 303–323.

Prünster, I. (2002). PhD in Mathematical Statistics. *Random probability measures derived from increasing additive processes and their application to Bayesian statistics.* Department of Mathematics, University of Pavia

Nieto-Barajas, L. E. and Walker, S. G. (2002). Markov Beta and Gamma processes for modelling hazard rates, *Scandinavian Journal of Statistics* 29(3): 413–424.

Regazzini, E., Lijoi, A. and Prünster, I. (2003). Distributional results for means of random measures with independent increments. *Ann. Statist.*, **31**, 560–585.

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

Walker, S.G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics*, **36**, 45–54.

Gracias!