Finite-dimensional discrete random structures and Bayesian clustering

Tommaso Rigon

University of Milano-Bicocca

Duke University - Statistical Science seminar

Outline

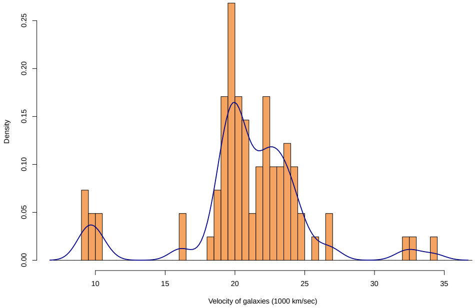
- Density estimation with truncated priors
- A Pitman–Yor multinomial process
- Normalized infinitely divisible processes
- An algorithm for discrete hierarchical processes (**maybe**!)

Joint work with

Antonio Lijoi
Bocconi University, Milano

Igor Prünster
Bocconi University, Milano

Mixtures for density estimation



- Let k be a kernel such that $\int_{\mathbb{R}^d} k(x, y)\, \mathrm{d}x = 1$ for any $y \in \Theta$
- Random probability measure $\tilde{p}_\infty = \sum_{j \geq 1} \pi_j \, \delta_{\theta_j}$ on Θ

$$\text{Random density} = \tilde{f}(x) = \int_\Theta k(x; y)\, \tilde{p}_\infty(\mathrm{d}y) = \sum_{j \geq 1} \pi_j k(x; \theta_j)$$

Mixtures of stick–breaking processes

Stick–breaking construction of $\tilde{p}$

- Pitman–Yor process: $\boxed{\tilde{p}_\infty \sim \mathrm{PY}(\sigma, \alpha; P_0)}$
- **Parameters** $\sigma \in [0, 1)$ and $\alpha > -\sigma$. Alternatively: $\sigma < 0$ and $\theta > -\alpha$.
- $(V_i)_{i \geq 1}$ such that $V_i \overset{\text{ind}}{\sim} \text{Beta}(1 - \sigma, \alpha + i\sigma)$.
- Probability weights & locations

$$\pi_1 = V_1, \qquad \pi_j = V_j \prod_{i=1}^{j-1}(1 - V_i) \quad j \geq 2, \qquad \theta_j \overset{\text{iid}}{\sim} P_0$$

  for some diffuse P_0 on Θ.
- $\tilde{p}_\infty \sim \mathrm{PY}(0, \alpha; P_0) \implies \tilde{p}_\infty$ is a **Dirichlet process** with parameters (α, P_0).

Inference on $\tilde{f}$?

- Sampling algorithms for Pitman–Yor mixture model include:
- **Marginal algorithms** integrate out $\tilde{p}_\infty$ for a Gibbs sampler that evaluates

$$\mathbb{E}[\,\tilde{f} \mid \text{data}\,].$$

- **Conditional algorithms**: truncate $\tilde{p}_\infty$ at a level H

$$\tilde{p}_{H,\text{tr}} = \sum_{j=1}^{H} \pi_j^* \, \delta_{\theta_j}, \quad \pi_1^* = \pi_1, \, \dots \, , \pi_{H-1}^* = \pi_{H-1}, \quad \pi_H^* = 1 - \sum_{i=1}^{H-1} \pi_i^*,$$

and simulate the posterior distribution $\tilde{p}_{H,\text{tr}} \mid \text{data}$.

Blocked Gibbs sampler

Ishwaran & James (2001): $\tilde{p}_\infty \mapsto \tilde{p}_{H,\mathrm{tr}}$

- Conditionally iid (latent) allocation variables $\nu_1, \dots, \nu_n$

$$\mathbb{P}(\nu_i = h \mid \tilde{p}_{H,\mathrm{tr}}) = \pi_h^*,$$

- Conditionally on $\boldsymbol{\nu} = (\nu_1, \dots, \nu_n)$, one has $y_i = \theta_{\nu_i}$ and

$$X_i \mid (\boldsymbol{\theta}, \boldsymbol{\nu}) \overset{\text{ind}}{\sim} k(X_i \mid \theta_{\nu_i}), \qquad i = 1, \dots, n.$$

- Distribution of $(\tilde{p}_{H,\mathrm{tr}} \mid -)$

$$\pi_1^* \mapsto V_1^*, \quad \pi_j^* \mapsto V_j^* \prod_{i=1}^{j-1}(1 - V_i^*),$$

  - $V_i^* \overset{\text{ind}}{\sim} \text{Beta}(1 - \sigma + n_i, \alpha + i\sigma + \sum_{h=i+1}^H n_h)$
  - $n_j = \text{card}\{i : \ \nu_i = j\}$

- Other full conditionals, i.e. $(\boldsymbol{\theta} \mid -)$ & $(\boldsymbol{\nu} \mid -)$, are easily identified

Blocked Gibbs sampler (ctd)

Popularity of the truncated stick–breaking approach

- It is **simple** to implement
- It **overcomes** some **"limitations" of Pólya urn**–type (or **marginal**) algorithms
  - problems when $k(x; \cdot)$ and P_0 are not conjugate
  - "underestimation" of uncertainty around point estimates $\mathbb{E}[\tilde{f} \mid \text{data}]$

However...

- **Not much is known** about $\tilde{p}_{H,\text{tr}}$
  - distribution of the random partition it induces
  - associated predictive distributions
  - ...

- Are there efficient alternatives, that work also for non–exchangeable data?
- Literature on priors on the finite-dimensional simplex is much more limited and mostly confined to the so–called **Dirichlet multinomial** process

Mixtures of NRMIs

- $(J_i, \theta_i)_{i \geq 1}$ are random points in $\mathbb{R}^+ \times \Theta$ such that $(J_i)_{i\geq 1} \perp (\theta_i)_{i \geq 1}$ and

$$\text{card}(\{(J_i, \theta_i) : \ i \geq 1\} \cap A) \sim \text{Po}(\nu(A)) \qquad \nu(A) = \alpha \int_A \rho(s)\, \mathrm{d}s\, P_0(\mathrm{d}\theta)$$

with P_0 a p.m. on Θ and $\rho : \mathbb{R}^+ \to \mathbb{R}^+$ such that $\int_0^\infty (s \wedge 1)\, \rho(s)\, \mathrm{d}s < +\infty$

- The resulting measure

$$\tilde{\mu} = \sum_{i \geq 1} J_i\, \delta_{\theta_i}$$

is **completely random**, i.e. $\tilde{\mu}(B) \perp \tilde{\mu}(B')$ if $B \cap B' = \varnothing$.

- **Normalized random measure** with independent increments (NRMI) are defined as

$$\pi_i = \frac{J_i}{\sum_{h \geq 1} J_h} \qquad \Longrightarrow \qquad \tilde{p}_\infty = \sum_{i \geq 1} \pi_i\, \delta_{\theta_i} \sim \text{NRMI}(\alpha, \rho; P_0)$$

under the additional condition $\int_0^\infty \rho(s)\, \mathrm{d}s = \infty$.

Truncated NRMIs

- **Marginal algorithms:** Integrate out $\tilde{p}_\infty$ and evaluate $\mathbb{E}[\tilde{f} \mid \text{data}]$ through a Gibbs sampler.
- **Conditional algorithms:**
  - Posterior characterization, with $\theta_1^*, \dots, \theta_k^*$ distinct values in $\boldsymbol{\theta}^{(n)} = (\theta_1, \dots, \theta_n)$

$$\tilde{p}_\infty \mid (\text{data}, \boldsymbol{\theta}^{(n)}) = \sum_{i \geq 1} \pi_{i,n}\, \delta_{\theta_i} + \sum_{j=1}^{k} \omega_{j,n}\, \delta_{\theta_j^*}$$

  - Representation due to Ferguson and Klass (1972)

$$\pi_{1,n} > \pi_{2,n} > \cdots$$

  - Simulate the truncated NRMI and re-normalize the weights

$$\boxed{\tilde{p}_{H,\text{tr}} = \sum_{i=1}^{H} \pi_{i,n}^*\, \delta_{\theta_i} + \sum_{j=1}^{k} \omega_{j,n}\, \delta_{\theta_j^*}}$$

- See Barrios et al. (2013) and Arbel & Prünster (2017).

Pros & cons

Pros

- **Ordering** of the $\pi_{i,n}$'s ensures that the truncation procedure retains most relevant random probability masses.
- Free from **limitations** of **Pólya urn**–type marginal algorithms.

Cons

- Simulation of the ordered weights $\pi^*_{i,n}$ in the truncated representation may lead to **numerical issues**.
- Not so much is known about the impact of this approximation on the posterior.

- Are there efficient alternatives, that work also for non–exchangeable data?
- For NRMIs the literature on priors on the finite-dimensional simplex is limited to the Dirichlet multinomial process.

The Dirichlet multinomial prior

- A classical finite-dimensional prior is the **Dirichlet multinomial** process

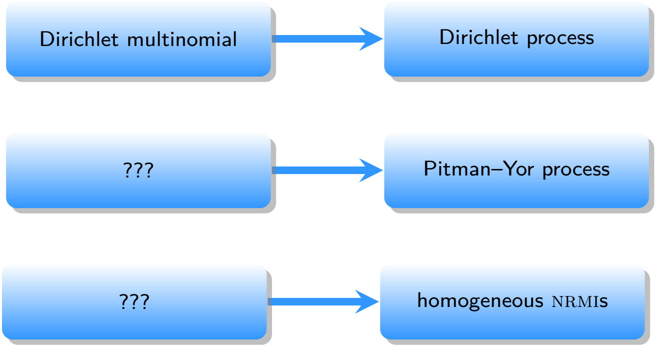
$$\tilde{p}_H = \sum_{h=1}^{H} \pi_h \delta_{\theta_h}, \quad (\pi_1, \dots, \pi_{H-1}) \sim \text{DIR}(\alpha/H, \dots, \alpha/H), \quad \theta_h \overset{\text{iid}}{\sim} P_0,$$

- The Dirichlet multinomial $\tilde{p}_H$ and the Dirichlet process $\tilde{p}_\infty \sim \text{PY}(0, \alpha, P_0)$ are related through

$$\mathcal{L}_{\tilde{p}_H} \Longrightarrow \mathcal{L}_{\tilde{p}_\infty}, \qquad H \to \infty.$$

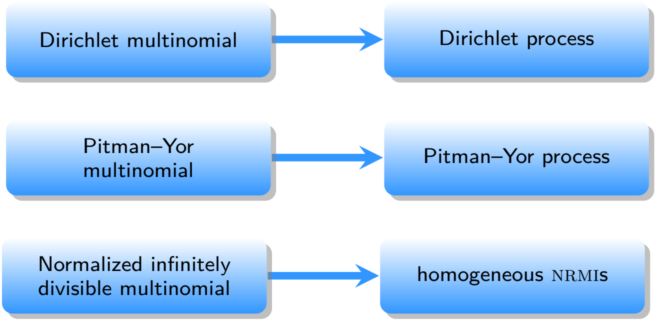
- The Dirichlet multinomial weakly converges to a Dirichlet process (Kingman, 1975).
- This motivates the usage of $\tilde{p}_H$ **also** as an approximation of $\tilde{p}_\infty$.
- However, the connection between $\tilde{p}_H$ and $\tilde{p}_\infty$ is far deeper, in fact holding for any finite value H.

Goal



- Natural extension of the Dirichlet multinomial?
- Would they preserve a good degree of analytical and computational tractability?

Goal



- Natural extension of the Dirichlet multinomial?
- Would they preserve a good degree of analytical and computational tractability? **Yes!**

The Pitman–Yor multinomial process

PY multinomial process

Let $\sigma \in [0,1)$ and $\alpha > -\sigma$. The random probability measure $\tilde{p}_H$ is a Pitman–Yor multinomial process if

$$(\tilde{p}_H \mid \tilde{p}_{0,H}) \sim \text{PY}(\sigma, \alpha; \tilde{p}_{0,H}), \qquad \tilde{p}_{0,H} = \frac{1}{H}\sum_{h=1}^{H} \delta_{\theta_h}, \qquad \theta_h \stackrel{\text{iid}}{\sim} P_0$$

In symbols $\tilde{p}_H \sim \text{PYM}_H(\sigma, \alpha; P_0)$.

- Discrete hierarchical Pitman–Yor process, with finitely supported base measure.
- Reminiscent of so–called discrete hierarchical processes (Camerlenghi et al., 2017)
- The baseline $\mathbb{E}[\tilde{p}_H \mid \tilde{p}_{0,H}] = \tilde{p}_{0,H}$ is **atomic**.
- This causes serious **analytical difficulties**, compared to the case where $\tilde{p}_{0,H}$ is non–atomic.

The PYM process: two alternative characterization

- In terms of **tempered stable** random variables:

$$J_h \mid U \sim \mathrm{TS}\left(\frac{1}{H}, \sigma; u\right) \quad \Rightarrow \quad \mathbb{E}\left(\mathrm{e}^{-\lambda J_h} \mid U = u\right) = \exp\left\{-\frac{1}{H}\left[(\lambda + u)^\sigma - u^\sigma\right]\right\},$$

$$U^\sigma \sim \mathrm{Ga}\left(\frac{\alpha}{\sigma}, 1\right).$$

If $\tilde{p}_H \sim \mathrm{PYM}_H(\sigma, \alpha; P_0)$, it can be equivalently represented as

$$\boxed{\tilde{p}_H = \sum_{h=1}^{H} \frac{J_h}{J^*}\,\delta_{\theta_h},} \qquad J^* = \sum_{h=1}^{H} J_h.$$

- In terms of **ratio–stable** random variables (Carlton, 2002). If $\tilde{p}_H \sim \mathrm{PYM}_H(\sigma, \alpha; P_0)$ and $(\pi_1, \dots, \pi_{H-1}) \sim \mathrm{RS}(\sigma, \alpha; 1/H, \dots, 1/H)$, then

$$\boxed{\tilde{p}_H = \sum_{h=1}^{H} \pi_h\,\delta_{\theta_h}}$$

Some remarks

- The **Dirichlet multinomial** is a special case: it corresponds to

$$(\pi_1, \ldots, \pi_{H-1}) \sim \mathrm{RS}(0, \alpha; 1/H, \ldots, 1/H)$$

- The density function of the weights $(\pi_1, \ldots, \pi_{H-1})$ is not available in closed form, the only exceptions being:

$$\mathrm{PYM}_H(1/2, 0; P_0) \quad \& \quad \mathrm{PYM}_H(0, \alpha; P_0).$$

- **Simulation** for Bayesian inference:
  - Simulate tempered stable random variables through an algorithm such as the one, e.g., in Ridout (2009)
  - numerical issues when $\sigma \approx 0$ if performed naïvely
    $\Rightarrow$ U tends to generate high values that cause overflows
    $\Rightarrow$ a suitable rescaling of J_h solves the issue

- Pitman & Yor (1997) $\Rightarrow$ J_h = polynomial tilting of a positive stable r.v.
  $\Rightarrow$ useful for posterior computations

NID multinomial processes

Normalized infinitely divisible multinomial processes

Let $\rho : \mathbb{R}^+ \to \mathbb{R}^+$ be such that $\int_0^\infty \rho(s)\,\mathrm{d}s = \infty$. The random probability measure $\tilde{p}_H$ is a normalized infinitely divisible multinomial process if

$$(\tilde{p}_H \mid \tilde{p}_{0,H}) \sim \mathrm{NRMI}(\alpha, \rho; \tilde{p}_{0,H}), \qquad \tilde{p}_{0,H} = \frac{1}{H}\sum_{h=1}^{H} \delta_{\theta_h}, \qquad \theta_h \overset{\text{iid}}{\sim} P_0$$

In symbols $\tilde{p}_H \sim \mathrm{NIDM}_H(\alpha, \rho; P_0)$.

- Same technical issues as those already outlined for the PYM process
- It can be described in terms of infinitely divisible random variables

$$\mathbb{E}\,\mathrm{e}^{-\lambda J} = \exp\left\{ -\alpha \int_0^\infty \left(1 - \mathrm{e}^{-\lambda s}\right) \rho(s)\,\mathrm{d}s \right\},$$

where $J \sim \mathrm{ID}(\alpha, \rho)$.

NID multinomial processes: a characterization

Let $\rho : \mathbb{R}^+ \to \mathbb{R}^+$ be such that $\int_0^\infty \rho(s)\,\mathrm{d}s = \infty$ and

$$J_h \overset{\text{iid}}{\sim} \mathrm{ID}\left(\frac{\alpha}{H}, \rho\right).$$

If $\tilde{p}_H \sim \mathrm{NIDM}_H(\alpha, \rho; P_0)$ then

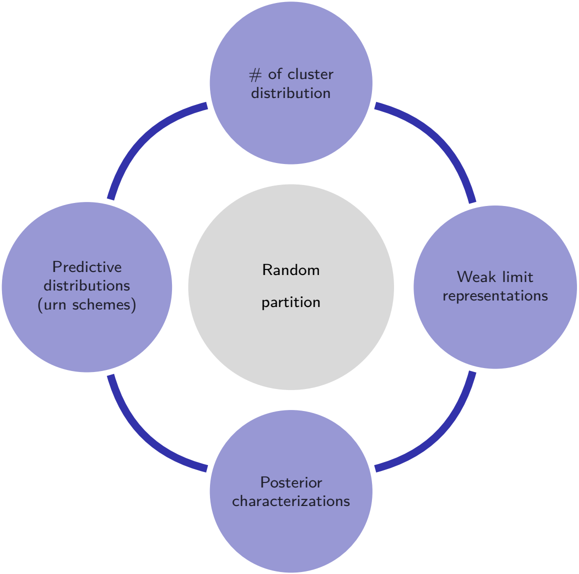
$$\tilde{p}_H = \sum_{h=1}^{H} \frac{J_h}{J^*} \delta_{\theta_h}, \qquad J^* = \sum_{h=1}^{H} J_h.$$

The vector of probability weights is referred to as **normalized infinitely divisible**

$$(\pi_1, \dots, \pi_{H-1}) = \left(\frac{J_1}{J^*}, \dots, \frac{J_{H-1}}{J^*}\right) \sim \mathrm{NID}\left(\frac{\alpha}{H}, \dots, \frac{\alpha}{H}; \rho\right)$$

- Dirichlet multinomial process $\Rightarrow$ $\rho(s) = s^{-1}\,\mathrm{e}^{-s}$
- Normalized inverse Gaussian multinomial process $\Rightarrow$ $\rho(s) = C\,s^{-3/2}\,\mathrm{e}^{-\tau s}$
- Generalized Dirichlet multinomial process $\Rightarrow$ $\rho(s) = s^{-1}\mathrm{e}^{-s}\sum_{j=0}^{\gamma}\mathrm{e}^{-js}$

Summary of the results



Exchangeable partition probability function (EPPF)

- Conditionally iid variables

$$Y_i \mid \tilde{p} \overset{\text{iid}}{\sim} \tilde{p}, \qquad \tilde{p} = \sum_{i \in \mathcal{I}} \pi_i \, \delta_{\theta_i},$$

  for some countable set $\mathcal{I}$.

- The discreteness of $\tilde{p}$ implies that with positive probability there will be ties among $Y_1, \ldots, Y_n$, therefore inducing a **random partition**, say Ψ_n, of $[n] = \{1. \ldots, n\}$

$$i \sim j \quad \Longleftrightarrow \quad Y_i = Y_j$$

- The probability distribution of Ψ_n is also known as **exchangeable partition probability function** (EPPF)

$$\Pi(n_1, \ldots, n_k) = \text{pr}(\Psi_n = \{C_1, \ldots, C_k\}) = \sum_{i_1 \neq \cdots \neq i_k} \mathbb{E}\left(\prod_{j=1}^{k} \pi_{i_j}^{n_j}\right),$$

  with $n_j = \text{card}(C_j)$ and $\sum_{j=1}^k n_j = n$.

- The EPPF **characterizes** $\tilde{p}$.

EPPFs (Lijoi, Prünster & R., 2020, 2023)

- The EPPF associated to a $\tilde{p}_H \sim \text{PYM}_H(\sigma, \alpha; P_0)$, with P_0 diffuse, is

$$\Pi_H(n_1, \ldots, n_k) = \frac{H!}{(H-k)!} \frac{1}{(\alpha+1)_{n-1}} \sum_{(\ell_1,\ldots,\ell_k)} \frac{\Gamma(\alpha/\sigma + |\boldsymbol{\ell}^{(k)}|)}{\sigma\Gamma(\alpha/\sigma+1)} \prod_{j=1}^k \frac{\mathscr{C}(n_j, \ell_j; \sigma)}{H^{\ell_j}},$$

with $\boldsymbol{\ell}^{(k)} = (\ell_1, \ldots, \ell_k) \in \times_{j=1}^k \{1, \ldots, n_j\}$, and $\mathscr{C}(n, k; \sigma)$ the GFC.

- The EPPF associated to a $\tilde{p}_H \sim \text{NIDM}_H(\alpha, \rho; P_0)$, with P_0 diffuse, is

$$\Pi_H(n_1, \ldots, n_k) = \frac{H!}{(H-k)!} \frac{1}{\Gamma(n)} \int_{\mathbb{R}^+} u^{n-1} \mathrm{e}^{-\alpha\psi(u)} \left\{ \prod_{j=1}^k V_{n_j}(u) \right\} \mathrm{d}u,$$

with $\psi(u) = \int_{\mathbb{R}^+} (1 - \mathrm{e}^{-us}) \rho(s) \, \mathrm{d}s$ and

$$V_m(u) = \left\{ (-1)^m \frac{\partial^m}{\partial u^m} \mathrm{e}^{-\frac{\alpha}{H}\psi(u)} \right\} \mathrm{e}^{\frac{\alpha}{H}\psi(u)}.$$

Distribution of $K_{n,H}$ (Lijoi, Prünster & R., 2020)

Once the EPPF $\Pi(n_1, \dots, n_k)$ is available, one can determine:

- The distribution of the **number of clusters**, or partitions sets, $K_{n,H}$.
- The **predictive distribution** of Y_{n+1}, conditional on $\boldsymbol{Y}^{(n)} = (Y_1, \dots, Y_n)$.
- The **posterior distribution** of $\tilde{p}$, given $\boldsymbol{Y}^{(n)}$.

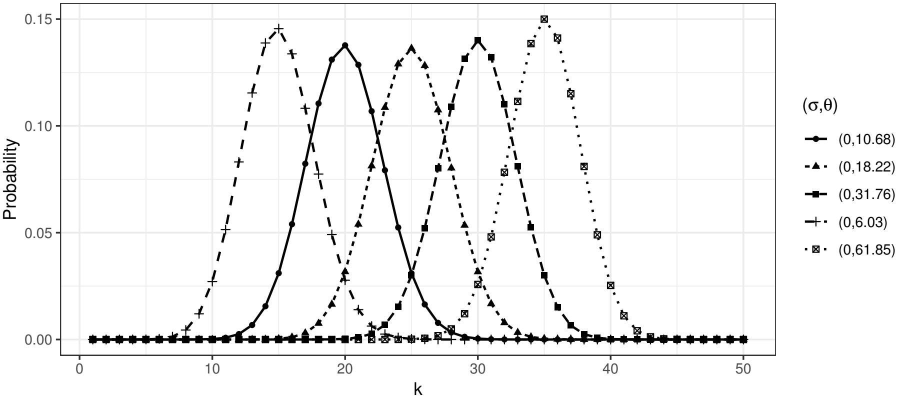
Example: distribution of $K_{n,H}$ of a PYM process

If $Y_1, \dots, Y_n \mid \tilde{p}_H \overset{\text{iid}}{\sim} \tilde{p}_H$ and $\tilde{p}_H \sim \text{PYM}(\sigma, \alpha; P_0)$, the number $K_{n,h}$ of distinct values in $\boldsymbol{Y}^{(n)}$ has probability distribution

$$\mathbb{P}(K_{n,H} = k) = \frac{H!}{(H-k)!} \frac{1}{(\alpha+1)_{n-1}} \sum_{\ell=k}^{n} \frac{(\alpha/\sigma + 1)_{\ell-1}}{\sigma H^\ell} \mathscr{S}(\ell, k) \mathscr{C}(n, \ell; \sigma)$$

$k \leq \min\{H, n\}$ and $\mathscr{S}(\ell, k)$ is the Stirling number of the second kind.

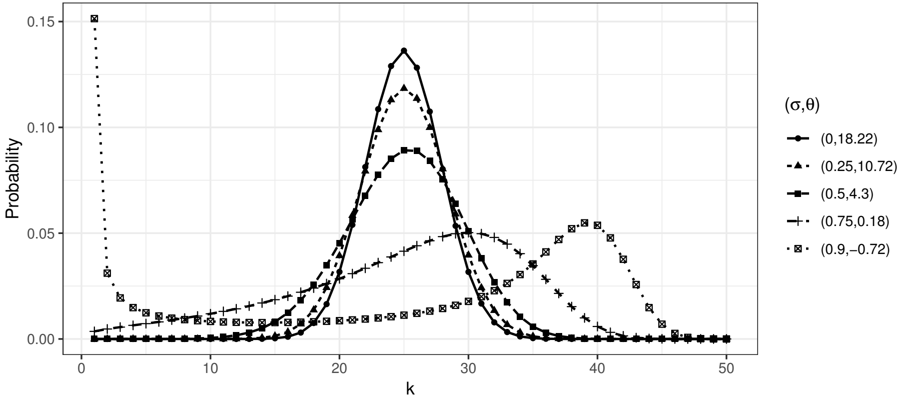
An example with the Dirichlet multinomial



- Distribution of the number of cluster when $n = 100$, $H = 50$ in the Dirichlet multinomial case ($\sigma = 0$).

- The Dirichlet multinomial is **highly informative**.

An example with the PYM, with $\sigma > 0$



- Distribution of the number of cluster when $n = 100$, $H = 50$, and for various choices of (σ, α) so that $\mathbb{E}(K_{n,H}) = 25$.

- More **robust** specification $\Longrightarrow$ De Blasi et al. (2015); Canale and Prünster (2017).

Urn schemes for the Pitman–Yor multinomial

- Urn-schemes for posterior inference in mixture models (Neal, 2000).
- The EPPF of a PYM can be interpreted as a **mixture** over

$$\mathbb{P}(\ell_1 = l_1, \dots, \ell_k = l_k \mid \boldsymbol{Y}^{(n)}) \propto \Gamma(\alpha/\sigma + |\boldsymbol{\ell}^{(k)}|) \prod_{j=1}^{k} \frac{\mathscr{C}(n_j, l_j; \sigma)}{H^{l_j}},$$

which can be efficiently and **independently** simulated, through data–augmentation.

Predictive distribution (Lijoi, Prünster & R., 2020)

Let $Y_1^*, \dots, Y_k^*$ be the distinct values recorded in $\boldsymbol{Y}^{(n)}$, then

$$\mathbb{P}(Y_{n+1} \in A \mid \boldsymbol{Y}^{(n)}) = \left(1 - \frac{k}{H}\right)\left(\frac{\alpha + |\bar{\boldsymbol{\ell}}^{(k)}|\sigma}{\alpha + n}\right) P_0(A) + \\ + \sum_{j=1}^{k} \left(\frac{1}{H}\frac{\alpha + |\bar{\boldsymbol{\ell}}^{(k)}|\sigma}{\alpha + n} + \frac{n_j - \bar{\ell}_j\sigma}{\alpha + n}\right)\delta_{Y_j^*}(A),$$

with $\bar{\boldsymbol{\ell}}^{(k)} = (\bar{\ell}_1, \dots, \bar{\ell}_k) = E(\boldsymbol{\ell}^{(k)} \mid \boldsymbol{Y}^{(n)})$.

Posterior of a PYM (Lijoi, Prünster & R., 2020)

- The **posterior distribution** of $\tilde{p}_H \sim \mathrm{PYM}_H(\sigma, \alpha; P_0)$, conditional on $\boldsymbol{Y}^{(n)}$ and $\boldsymbol{\ell}^{(k)}$, is

$$(\tilde{p}_H \mid \boldsymbol{Y}^{(n)}, \boldsymbol{\ell}^{(k)}) = \sum_{j=1}^{k} (W_j + W_{k+1} R_j)\delta_{Y_j^*} + W_{k+1} \sum_{j=k+1}^{H} R_j \delta_{\theta_j},$$

  where $\theta_j \overset{\text{iid}}{\sim} P_0$. Moreover,

$$(W_1, \dots, W_k \mid \boldsymbol{Y}^{(n)}, \boldsymbol{\ell}^{(k)}) \sim \mathrm{DIR}(n_1 - \ell_1\sigma, \dots, n_k - \ell_k\sigma, \alpha + |\boldsymbol{\ell}^{(k)}|\sigma)$$

$$(R_1, \dots, R_{H-1} \mid \boldsymbol{Y}^{(n)}, \boldsymbol{\ell}^{(k)}) \sim \mathrm{RS}(\sigma, \alpha + |\boldsymbol{\ell}^{(k)}|\sigma; 1/H, \dots, 1/H)$$

  are independent vectors.

- The PYM might be referred to as being **quasi-conjugate**.

- It resembles the infinite–dimensional structure (i.e. the Pitman–Yor)

$$\tilde{p}_\infty \mid \boldsymbol{Y}^{(n)} = \sum_{j=1}^{k} W_j' \, \delta_{Y_j^*} + W_{k+1}' \, \mathrm{PY}(\sigma, \alpha + k\sigma; P_0)$$

  and $(W_1', \dots, W_k') \sim \mathrm{DIR}(n_1 - \sigma, \dots, n_k - \sigma, \alpha + k\sigma)$.

Urn schemes for NIDM processes

- Let us define

$$\Delta_{m,H}(u) = \sum_{\ell=1}^{m} \left(\frac{\theta}{H}\right)^{\ell-1} \frac{1}{\ell!} \sum_{\boldsymbol{q}} \binom{m}{q_1 \cdots q_\ell} \prod_{r=1}^{\ell} \int_0^\infty s^{q_r} e^{-us} \rho(s)\mathrm{d}s,$$

over all vectors $\boldsymbol{q} = (q_1, \dots, q_\ell)$ of positive integers s.t. $\sum_{r=1}^{\ell} q_r = m$.

- The density of the **latent** random variable $(U_{n,H} \mid X^{(n)})$ is

$$f_H(u \mid \boldsymbol{Y}^{(n)}) \propto u^{n-1} e^{-c\psi(u)} \prod_{j=1}^{k} \Delta_{n_j,H}(u).$$

Predictive distribution (Lijoi, Prünster & R., 2023)

$$\mathbb{P}(Y_{n+1} \in A \mid \boldsymbol{Y}^{(n)}) = \left(1 - \frac{k}{H}\right)\frac{\theta}{n} \int_{\mathbb{R}^+} u\Delta_{1,H}(u) f_H(u \mid \boldsymbol{Y}^{(n)})\mathrm{d}u\, P_0(A) + \\ + \sum_{j=1}^{k} \frac{1}{n} \int_{\mathbb{R}^+} u \frac{\Delta_{n_j+1,H}(u)}{\Delta_{n_j,H}(u)} f_H(u \mid \boldsymbol{Y}^{(n)})\mathrm{d}u\, \delta_{Y_j^*}(A).$$

Posterior characterization of a NIDM

- The **posterior distribution** of $\tilde{p}_H = \tilde{\mu}_H/\tilde{\mu}(\mathbb{X}) \sim \text{NIDM}(\theta, \rho; P)$, conditional on $\boldsymbol{Y}^{(n)}$, is a mixture with respect to $U_{n,H}$, with

$$(\tilde{\mu}_H \mid \boldsymbol{Y}^{(n)}, U_{n,H}) \stackrel{d}{=} \sum_{j=k+1}^{H} J_j \delta_{\theta_j} + \sum_{j=1}^{k} (J_j + J_j^{(u)}) \delta_{Y_j^*},$$

where $\theta_{k+1}, \dots, \theta_H$ are iid draws from P_0 and

(i) *the jumps* $(J_h \mid \boldsymbol{Y}^{(n)}, U_{n,H})$ *for* $h = 1, \dots, H$ *are iid* $\text{ID}\left(\theta/H, \rho^{(u)}\right)$ *r.v. with*

$$\rho^{(u)}(s) = e^{-Us}\rho(s)$$

(ii) *the jumps* $(J_j^{(u)} \mid \boldsymbol{Y}^{(n)}, U_{n,H})$ *for* $j = 1, \dots, K_n$ *are independent and nonnegative r.v. such that*

$$\mathbb{E}\left(e^{-\lambda J_j^{(u)}} \,\middle|\, \boldsymbol{Y}^{(n)}, U\right) = \Delta_{n_j,H}(\lambda + U)/\Delta_{n_j,H}(U)$$

(iii) $(J_h \mid \boldsymbol{Y}^{(n)}, U_{n,H})$ *and* $(J_j^{(u)} \mid \boldsymbol{Y}^{(n)}, U_{n,H})$ *are mutually independent.*

Weak limits, as $H \nearrow \infty$

Weak limit for PYM processes (Lijoi, Prünster & R., 2020)

Let $\tilde{p}_H \sim \text{PYM}_H(\sigma, \alpha; P_0)$ and $\tilde{p}_\infty \sim \text{PY}(\sigma, \theta; P_0)$. Then

$$\mathcal{L}_{\tilde{p}_H} \Longrightarrow \mathcal{L}_{\tilde{p}_\infty} \qquad H \nearrow \infty$$

Weak limit for NIDM processes (Lijoi, Prünster & R., 2023)

Let $\tilde{p}_H \sim \text{NIDM}_H(\alpha, \rho; P_0)$ and $\tilde{p}_\infty \sim \text{NRMI}(\alpha, \rho; P_0)$. Then

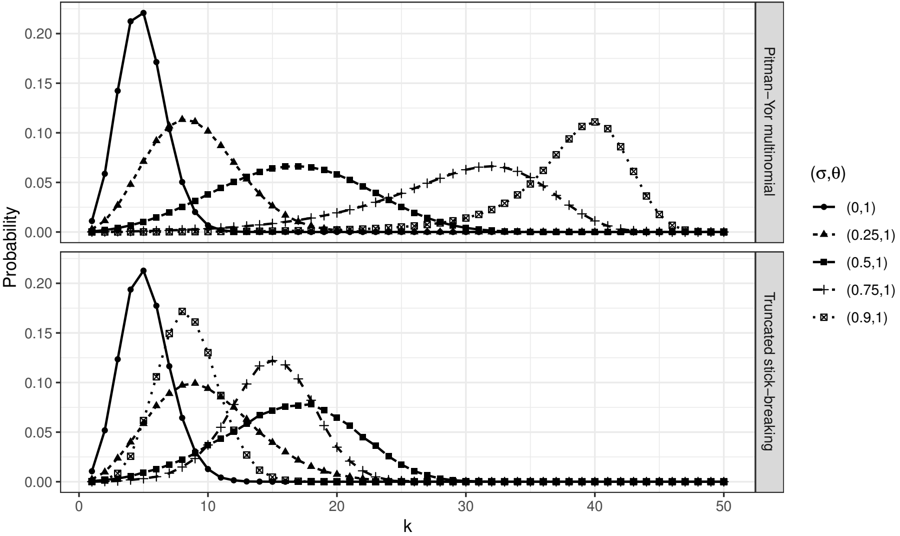
$$\mathcal{L}_{\tilde{p}_H} \Longrightarrow \mathcal{L}_{\tilde{p}_\infty} \qquad H \nearrow \infty$$

- In practice, one may select the approximation level H by inspecting the EPPFs or the cluster distributions: bounds for

$$\Pi_H(n_1, \dots, n_k)/\Pi_\infty(n_1, \dots, n_k) \quad \& \quad d_{\text{TV}}(m_H, m_\infty)$$

- Distributional properties hold for any choice of H, not just at the limit.
- This is not true for the **truncated stick-breaking** construction. **Problems** if the truncation is chosen lightly.

The distribution of the number of clusters: $\tilde{p}_H$ vs $\tilde{p}_{H,\mathrm{tr}}$



Non–exchangeable data

Observations from different groups/samples

$$\boldsymbol{X}_1^{(n_1)} = (X_{1,1}, \dots, X_{1,n_1}), \; \dots, \; \boldsymbol{X}_d^{(n_d)} = (X_{d,1}, \dots, X_{d,n_d})$$

- Data from different, though related experiments: multi–center studies, change–point problems, meta–analysis, ...
- Exchangeability within each group, not across groups
- **Partial exchangeability:** at the core of modern approaches (latent Dirichlet allocation for topics modeling, infinite hidden Markov models, etc.).

Conditional independence

$$\begin{aligned} X_{i,\kappa} \mid (\tilde{p}_1, \dots, \tilde{p}_d) &\sim \tilde{p}_i \qquad\qquad \kappa = 1, \dots, n_i \\ X_{i,\kappa}, X_{j,\ell} \mid (\tilde{p}_1, \dots, \tilde{p}_d) &\sim \tilde{p}_i \times \tilde{p}_j \\ (\tilde{p}_1, \dots, \tilde{p}_d) &\sim Q_d \end{aligned}$$

Here we will consider

$$Q_d = \text{ law of a hierarchical process}$$

The hierarchical NRMI-PY process

Let $X_{i,\kappa}$ for $\kappa = 1, \ldots, n_i$ and $i = 1, \ldots, d$, be an array of d collections of random variables such that

$$(X_{i,\kappa} \mid \tilde{p}_i) \overset{\text{ind}}{\sim} \tilde{p}_i \qquad \kappa = 1, \ldots, n_i, \quad i = 1, \ldots, d$$
$$(\tilde{p}_i \mid \tilde{p}_0) \overset{\text{iid}}{\sim} \text{NRMI}(\alpha, \rho, \tilde{p}_0), \qquad i = 1, \ldots, d$$
$$\tilde{p}_0 \sim \text{PY}(\sigma_0, \alpha_0, P_0)$$

where P_0 is a **diffuse** probability measure defined on $\mathbb{X}$.

$$\tilde{p}_i = \sum_{j \geq 1} \pi_{j,i}\, \delta_{\theta_{i,j}}, \qquad \theta_{i,j} | \tilde{p}_0 \overset{\text{iid}}{\sim} \tilde{p}_0, \qquad \tilde{p}_0 = \sum_{j \geq 1} \pi_{j,0}\, \delta_{\theta_{0,j}}$$

- General theory in Camerlenghi et al. (2019)
- The hierarchical Dirichlet process of Teh et al. (2006) is a particular case.

Alternative representation

Definition of NIDP's

- Let $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots)$ be a collection of non–negative numbers s.t. $\sum_{h=1}^{\infty} \alpha_h < \infty$.
- An infinite random vector $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots)$ such that $\sum_{h=1}^{\infty} \pi_h = 1$ a.s. is termed ***normalized infinitely divisible process*** if, for finite partition $\mathcal{H}_1, \dots, \mathcal{H}_M$ of $\mathbb{N}$, one has

$$\left(\sum_{j \in \mathcal{H}_1} \pi_j, \dots, \sum_{j \in \mathcal{H}_{M-1}} \pi_j\right) \sim \mathrm{NID}\left(\sum_{j \in \mathcal{H}_1} \alpha_j, \dots, \sum_{j \in \mathcal{H}_M} \alpha_j; \rho\right),$$

and it will be denoted $\boldsymbol{\pi} \sim \mathrm{NIDP}(\boldsymbol{\alpha}, \rho)$.

- If $\tilde{p}_0 = \sum_{j \geq 1} \pi_{j,0}\, \delta_{\theta_{j,0}} \sim \mathrm{PY}(\sigma_0, \alpha_0; P_0)$, then

$$\boldsymbol{\pi}_0 = (\pi_{1,0}, \pi_{2,0}, \dots) \sim \mathrm{GEM}(\sigma_0, \alpha_0)$$

- Allocation variables $G_{i,\kappa}$

$$G_{i,\kappa} \mapsto Y_{i,\kappa} = \theta_{G_{i,\kappa},0}$$

Alternative representation (ctd)

NRMI–PY hierarchical process

The NRMI-PY model can be expressed in an equivalent manner

$$\boldsymbol{\pi}_0 \sim \text{GEM}(\sigma_0, \alpha_0), \qquad \theta_{j,0} \overset{\text{iid}}{\sim} P_0,$$

$$(\boldsymbol{\pi}_i \mid \boldsymbol{\pi}_0) \overset{\text{iid}}{\sim} \text{NIDP}(\alpha\boldsymbol{\pi}_0, \rho), \qquad (G_{i,\kappa} \mid \boldsymbol{\pi}_i) \overset{\text{iid}}{\sim} \text{CATEGORICAL}(\boldsymbol{\pi}_i),$$

for $\kappa = 1, \dots, n_i$ and $i = 1, \dots, d$.

- Key idea: **truncating the GEM sequence up to the *H*-th term has a cascade effect**, because the infinite dimensional NIDP reduces to a NID
- With a truncation on $\tilde{p}_0$, we end up working with multinomial processes $\tilde{p}_{i,H}$ at the level of the single populations
- Almost sure upper bounds for $\mathrm{d}_{\text{TV}}(\tilde{p}_i, \tilde{p}_{i,H})$
- Cases where the NID vectors $(\pi_{i,1}, \dots, \pi_{i,H-1})$ have **distributions** available in **closed form**.

From discrete priors to mixture models

Continuous data: $X_{i,\kappa}$ for $\kappa = 1, \dots, n_i$ and $i = 1, \dots, d$

Let $k : \mathbb{X} \times \Theta \to \mathbb{R}^+$ a transition kernel such that for any $y \in \Theta$

$$x \mapsto k(x; y) \qquad \text{is a density}$$

Model:

$$\boldsymbol{\pi}_0 \sim \mathrm{GEM}_H(\sigma_0, \alpha_0,), \qquad \theta_{h,0} \overset{\text{iid}}{\sim} P_0,$$

$$(\boldsymbol{\pi}_i \mid \boldsymbol{\pi}_0) \overset{\text{iid}}{\sim} \mathrm{NID}(\alpha\pi_{1,0}, \dots, \alpha\pi_{H,0}; \rho), \qquad (G_{i,\kappa} \mid \boldsymbol{\pi}_i) \overset{\text{iid}}{\sim} \mathrm{CAT}(\boldsymbol{\pi}_i),$$

$$(X_{i,\kappa} \mid G_{i,\kappa}, \boldsymbol{\theta}_0) \overset{\text{ind}}{\sim} k(x; \theta_{G_{i,\kappa},0}),$$

- Marginalizing over the allocation variables $G_{i,\kappa}$, we obtain a finite mixture model

$$(Y_{i,\kappa} \mid \boldsymbol{\pi}_i, \boldsymbol{\theta}_0) \overset{\text{ind}}{\sim} f_i(y \mid \boldsymbol{\pi}_i, \boldsymbol{\theta}_0) = \sum_{h=1}^{H} \pi_{i,h} k(y; \theta_{h,0}).$$

Gibbs sampler

for *i from* 1 *to d* **do** Sample the cluster indicator $G_{i,\kappa}$ independently

for κ *from* 1 *to* n_i **do**

$$\mathbb{P}(G_{i,\kappa} = h \mid -) = \frac{\pi_{i,h}\, k(X_{i,\kappa};\theta_{h,0})}{\sum_{h'=1}^{H} \pi_{i,h'}\, k(X_{i,\kappa};\theta_{h',0})}, \qquad h = 1,\dots,H.$$

for *i from* 1 *to d* **do** Sample $\boldsymbol{\pi}_i$ independently from the full conditional

$$f(\boldsymbol{\pi}_i \mid -) \propto f(\boldsymbol{\pi}_i \mid \boldsymbol{\pi}_0) \prod_{h=1}^{H} \pi_{i,h}^{n_{i,h}}, \qquad n_{i,h} = \sum_{j=1}^{n_i} \mathbb{1}(G_{i,\kappa} = h).$$

Sample the baseline mixing parameter $\boldsymbol{\pi}_0$

$$f(\boldsymbol{\pi}_0 \mid -) \propto f(\boldsymbol{\pi}_0) \prod_{i=1}^{d} f(\boldsymbol{\pi}_i \mid \boldsymbol{\pi}_0).$$

for *h from* 1 *to H* **do** Sample the kernel parameters $Z_{h,0}$ independently

$$f(\theta_{h,0} \mid -) \propto f(\theta_{h,0}) \prod_{(i,\kappa)\in\mathcal{G}_h} k(X_{i,\kappa};\theta_{h,0}), \quad \mathcal{G}_h = \{i,\kappa : G_{i,\kappa} = h\}.$$

Concluding remarks

- **Multinomial processes** have appealing properties, and they are both analytically and computationally tractable.
- Their use may ease the implementation of BNP procedures with more complex dependence structures than exchangeability.
- Future work might include the implementation of these prior for **species discovery**.
- **Hierarchical processes**: the key idea of our approach are the approximation of the baseline $\tilde{p}_0$ and, then, use of the resulting multinomial processes at the level of the single groups/samples.
- Truncation of $\tilde{p}_0$ and be either the stick-breaking or, again, multinomial approximations.

Grazie!

Lijoi, Prünster & Rigon (2020). The Pitman–Yor multinomial process for mixture modeling. *Biometrika*, **107**(4), 891–906

Lijoi, Prünster & Rigon (2020). Sampling hierarchies of discrete random structures. *Statistics & Computing*, **30**, 1591–1607

Lijoi, Prünster & Rigon (2023). Finite-dimensional discrete random structures and Bayesian clustering. *Journal of the American Statistical Association (T&M)*, in press