Multivariate Species Sampling Process (mSSP)

Beatrice Franzolini
(Joint work with Lijoi, Prünster and Rebaudo)

Main References:

- Franzolini, B., Lijoi, A., Prünster, I., & Rebaudo, G. (2025). Multivariate species sampling models. arXiv preprint arXiv:2503.24004.
- Franzolini, B., Lijoi, A., Prünster, I., & Rebaudo, G. (2025+). Partially exchangeable random partition structures (Ongoing).

Università Bocconi
BIDSA
Bocconi Institute
for Data Science and Analytics

Species sampling processes (Pitman, 1996; Balocchi, Favaro & Naulet, 2024)

(Univariate) Species Sampling Process

A random probability $\tilde{P}$ is a species sampling process if

$$\tilde{P} \overset{a.s.}{=} \sum_{h \geq 1} \pi_h \delta_{\theta_h} + (1 - \sum_{h \geq 1} \pi_h) P_0$$

- where $(\theta_h)_{h\geq 1} \perp\!\!\!\perp (\pi_h)_{h\geq 1}$
- $\theta_h \overset{\text{iid}}{\sim} P_0$ with P_0 non–atomic
- the probability weights $(\pi_h)_{h\geq 1} \sim \mathcal{L}_\pi$ are such that $\sum_{h\geq 1} \pi_h \leq 1$ a.s.
  We write $\tilde{P} \sim \boldsymbol{SSP(\mathcal{L}_\pi, P_0)}$.

Species sampling processes (Pitman, 1996; Balocchi, Favaro & Naulet, 2024)

(Univariate) Species Sampling Process

A random probability $\tilde{P}$ is a species sampling process if

$$\tilde{P} \overset{a.s.}{=} \sum_{h \geq 1} \pi_h \delta_{\theta_h} + (1 - \sum_{h \geq 1} \pi_h) P_0$$

- where $(\theta_h)_{h \geq 1} \perp\!\!\!\perp (\pi_h)_{h \geq 1}$
- $\theta_h \overset{\text{iid}}{\sim} P_0$ with P_0 non–atomic
- the probability weights $(\pi_h)_{h \geq 1} \sim \mathcal{L}_\pi$ are such that $\sum_{h \geq 1} \pi_h \leq 1$ a.s.

  We write $\tilde{P} \sim \boldsymbol{SSP(\mathcal{L}_\pi, P_0)}$.

SSP in Bayesian statistics

- Used for **species sampling problems**, which involve unknown discrete distributions.
- Used in mixture models with likelihood $p(y \mid \tilde{P}) = \int_{\mathbb{X}} k(y; x) \tilde{P}(\mathrm{d}x)$
  - for **density estimation**
  - for **model-based clustering**

  They include all most used mixture models (finite mixtures, mixtures of finite mixtures, infinite mixtures)

Species sampling processes (Pitman, 1996; Balocchi, Favaro & Naulet, 2024)

(Univariate) Species Sampling Process

A random probability $\tilde{P}$ is a species sampling process if

$$\tilde{P} \overset{a.s.}{=} \sum_{h\geq 1} \pi_h \delta_{\theta_h} + (1 - \sum_{h\geq 1} \pi_h) P_0$$

- where $(\theta_h)_{h\geq 1} \perp\!\!\!\perp (\pi_h)_{h\geq 1}$
- $\theta_h \overset{\text{iid}}{\sim} P_0$ with P_0 non–atomic
- the probability weights $(\pi_h)_{h\geq 1} \sim \mathcal{L}_\pi$ are such that $\sum_{h\geq 1} \pi_h \leq 1$ a.s.

  We write $\tilde{P} \sim \boldsymbol{SSP(\mathcal{L}_\pi, P_0)}$.

SSP in Bayesian statistics

- Used for **species sampling problems**, which involve unknown discrete distributions.
- Used in mixture models with likelihood $p(y \mid \tilde{P}) = \int_{\mathbb{X}} k(y; x) \tilde{P}(\mathrm{d}x)$
  - for **density estimation**
  - for **model-based clustering**

  They include all most used mixture models (finite mixtures, mixtures of finite mixtures, infinite mixtures)

More importantly: they are a structural class

SSPs are in one-to-one correspondence with **partitions of (infinitely) exchangeable objects**

Exchangeable partitions and SSP

More importantly:
SSPs are in one-to-one correspondence with **partitions of (infinitely) exchangeable objects**

Exchangeable partitions and SSP

More importantly:
SSPs are in one-to-one correspondence with **partitions of (infinitely) exchangeable objects**

Exchangeable partition

A collection of random partitions $\Pi = (\Pi_n)_{n \geq 1}$,
where, Π_n is a partition of $\{1, 2, \dots, n\}$, for every $n \in \mathbb{N}$,
is **exchangeable**, if

- Π_n can be obtained eliminating the element $n+1$ from Π_{n+1}
- for any permutation σ of n elements,[a] $\mathbb{P}(\Pi_n = p_n) = \mathbb{P}(\Pi_n = \sigma(p_n))$

We write $\Pi \sim$ EPPF.

[a] $\sigma(p_n)$ denotes the partition obtained permuting the elements in the sets of p_n accordingly to σ

E.g., $n = 5$, and, $\sigma = (1, 5, 4, 2)$

$$\mathbb{P}\left(\Pi_5 = \{3, 1, 4\}\ \{5, 2\}\right) = \mathbb{P}\left(\Pi_5 = \{3, 5, 2\}\ \{4, 1\}\right)$$

Exchangeable partitions, SSP, and species sampling sequences

More importantly:
SSPs are in one-to-one correspondence with **exchangeable partitions**

Sampling observations $(X_1, \ldots, X_n)$, for any n, via the hierarchical model

$$X_i \mid \tilde{P} \overset{iid}{\sim} \tilde{P} \qquad \tilde{P} \sim SSP(\mathcal{L}_\pi, P_0)$$

or

step. 1 sampling an exchangeable random partition Π_n,
step. 2 independently sampling unique values $\{X_1^*, \ldots, X_K^*\}$ for each set from a non-atomic prob. measure P_0.

X_2 X_{15}
X_1 X_{12}
X_5
X_{13} X_{14} X_{11}
X_3
X_7
X_4
X_9
X_{10} X_8 X_6

produces the same law for the exchangeable sequence $(X_1, \ldots, X_n, \ldots)$, called species sampling sequence.

Exchangeable partitions, SSP, and species sampling sequences

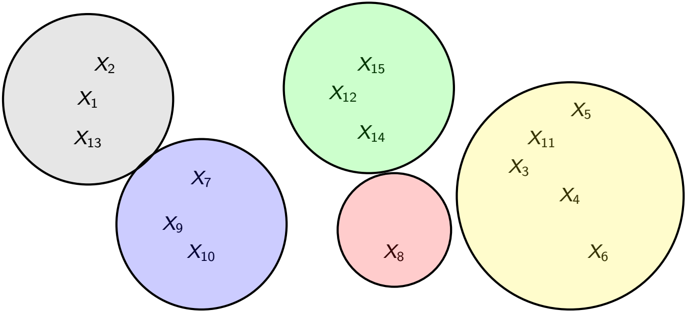
More importantly:
SSPs are in one-to-one correspondence with **exchangeable partitions**

Sampling observations $(X_1, \ldots, X_n)$, for any n, via the hierarchical model

$$X_i \mid \tilde{P} \overset{iid}{\sim} \tilde{P} \qquad \tilde{P} \sim SSP(\mathcal{L}_\pi, P_0)$$

or

step. 1 sampling an exchangeable random partition Π_n,
step. 2 independently sampling unique values $\{X_1^*, \ldots, X_K^*\}$ for each set from a non-atomic prob. measure P_0.



produces the same law for the exchangeable sequence $(X_1, \ldots, X_n, \ldots)$, called species sampling sequence.

Exchangeable partitions, SSP, and species sampling sequences

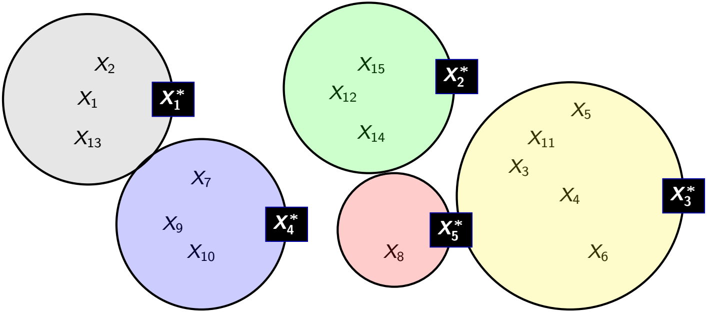
More importantly:
SSPs are in one-to-one correspondence with **exchangeable partitions**

Sampling observations $(X_1, \dots, X_n)$, for any n, via the hierarchical model

$$X_i \mid \tilde{P} \overset{iid}{\sim} \tilde{P} \qquad \tilde{P} \sim SSP(\mathcal{L}_\pi, P_0)$$

or

step. 1 sampling an exchangeable random partition Π_n,
step. 2 independently sampling unique values $\{X_1^*, \dots, X_K^*\}$ for each set from a non-atomic prob. measure P_0.



produces the same law for the exchangeable sequence $(X_1, \dots, X_n, \dots)$, called species sampling sequence.

Why do we need a multivariate version?

- Because SSPs are appropriate **only** to describe and model exchangeable sequences / data $(X_1, \dots, X_n, \dots)$.
- Because many discrete processes for data beyond exchangeability have been proposed, but no unifying class has been studied, leaving many open questions.

Partial exchangeability

Let us consider

- The observable sequence $\boldsymbol{X} = (X_1, X_2, \dots, X_n, \dots)$
- Additional information: group division $\boldsymbol{d} = (d_1, \dots, d_n, \dots,)$, with $d_i \in \{1, \dots, J\}$

$\boldsymbol{X}$ is **partially exchangeable** with respect to $\boldsymbol{d}$ if for any $n \geq 1$ and for $\boldsymbol{d}$-invariant permutation σ, i.e., $d_{\sigma(i)} = d_i$,

$$(X_1, \dots, X_n) \stackrel{d}{=} (X_{\sigma(1)}, \dots, X_{\sigma(n)})$$

Theorem (de Finetti, 1938)

if $\sum_{i=1}^{+\infty} \mathbb{1}(d_i = j) = \infty, \forall j,$

$\boldsymbol{X}$ is **partially exchangeable** with respect to $\boldsymbol{d}$, if and only if

$$X_i \mid \tilde{P}_1, \dots, \tilde{P}_J \stackrel{ind}{\sim} \tilde{P}_{d_i} \qquad \text{for } i = 1, \dots, n$$
$$(\tilde{P}_1, \dots, \tilde{P}_J) \sim Q$$

Dependent nonparametric priors

1. Additive structures
   - First proposed by Müller, Quintana & Rosner (2004) for the Dirichlet process:
   $$\tilde{P}_j = \epsilon_j Q_0 + (1-\epsilon_j) Q_j, \quad Q_j \overset{ind}{\sim} \mathrm{DP}(\alpha_j, P_0).$$
   - For general normalized random measures by Lijoi, Nipoti & Prunster (2014).
2. Hierarchical structures
   - First proposed by Teh, Jordan, Beal & Blei (2006) for the Dirichlet process: Hierarchical Dirichlet process (HDP).
   $$\tilde{P}_j \mid Q \overset{iid}{\sim} \mathrm{DP}(\alpha, Q), \quad Q \sim \mathrm{DP}(\alpha_0, P_0).$$
   - Generalization of HDP includes HPYP (Teh, 2006; Battiston, Favaro & Teh, 2018; Camerlenghi, Lijoi, Orbanz & Prünster, 2019), HNCRM (Camerlenghi, Lijoi, Orbanz & Prünster, 2019; Argiento, Cremaschi & Vannucci, 2020), HSSP (Bassetti, Casarin & Rossini, 2020).
3. Nested structures
   - First proposed by Rodriguez, Dunson & Gelfand (2008) for the Dirichlet process with stick-breaking representation: Nested Dirichlet process (NDP)
   $$\tilde{P}_j \mid Q \overset{iid}{\sim} Q, \quad Q \sim \mathrm{DP}(\alpha, \mathrm{DP}(\beta, P_0)).$$
   - Generalization to Nested NCRM and PYP (Camerlenghi, Dunson, Lijoi, Prünster & Rodriguez, 2019).
4. Composition of the previous classes: semi-HDP (Beraha, Guglielmi & Quintana, 2021), HHDP (Lijoi, Prünster & Rebaudo, 2023), nCAM (Denti, Camerlenghi, Guindani & Mira, 2023).
5. Other single-atoms dependent processes (MacEachern, 1999, 2000; Quintana, Müller, Jara & MacEachern, 2022): tree stick-breaking (Horiguchi et al., 2022), Compound random measures (Griffin & Leisen, 2017), vectors of normalized independent finite point processes (Colombi et al., 2023).
6. Normalized completely random vector (Catalano, Lijoi & Prünster, 2021).

Correlation as measure of dependence

Correlation

The most popular measure of dependence is correlation: since typically

$$\mathrm{corr}(\tilde{P}_1(A), \tilde{P}_2(A)),$$

does not depend on A it is taken as a measure of overall dependence.
Statistical implication: borrowing of information!

Examples:

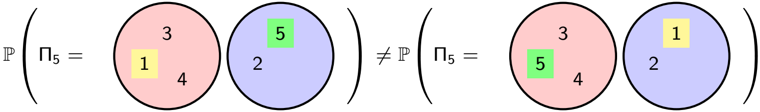
- Hierarchical Dirichlet process (HDP)
$$\mathrm{corr}(\tilde{P}_1(A), \tilde{P}_2(A)) = \frac{c+1}{c+1+c_0}$$
- Nested Common Atom model (n-CAM)
$$\mathrm{corr}(\tilde{P}_1(A), \tilde{P}_2(A)) = 1 - \frac{\beta\alpha}{(2\beta+1)(1+\alpha)}$$

Partially exchangeable partitions

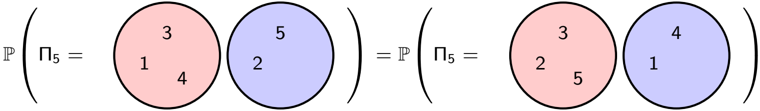
Given a group division $\boldsymbol{d} = (d_1, \ldots, d_n, \ldots)$ with $d_i \in \{1, \ldots, J\}$,

> *Partially exchangeable partition*
>
> A collection of random partitions $\Pi = (\Pi_n)_{n \geq 1}$,
> where, Π_n is a partition of $\{1, 2, \ldots, n\}$, for every $n \in \mathbb{N}$,
> is **partially exchangeable with respect to $\boldsymbol{d}$**, if
>
> - Π_n can be obtained eliminating the element $n+1$ from Π_{n+1}
> - for any $\boldsymbol{d}$-invariant permutation σ of n elements, $\mathbb{P}(\Pi_n = p_n) = \mathbb{P}(\Pi_n = \sigma(p_n))$
>
> We write $\Pi \sim$ pEPPF.

E.g., $n = 5$, $d = (1, 1, 1, 2, 2)$, and $\sigma = (1, 5)$



E.g., $n = 5$, $d = (1, 1, 1, 2, 2)$, and $\sigma = (1, 2)(4, 5)$



Multivariate species sampling processes

Given the sampling mechanism of the Bayesian model

$$X_i \mid (\tilde{P}_1, \dots, \tilde{P}_J) \stackrel{ind}{\sim} \tilde{P}_{d_i} \qquad (\tilde{P}_1, \dots, \tilde{P}_J) \sim Q \tag{1}$$

Definition mSSP

$(\tilde{P}_1, \dots, \tilde{P}_J)$ is a **multivariate species sampling process (mSSP)**,
if sampling $\boldsymbol{X}$ from (1) is equivalent to

- step. 1 sampling a **partially exchangeable random partition** Π_n,
- step. 2 independently sampling unique values $\{X_1^*, \dots, X_K^*\}$ for each set from a non-atomic prob. measure P_0.

$X_2, d_2 = 1$ $X_{15}, d_{15} = 1$

$X_1, d_1 = 1$ $X_{12}, d_{12} = 2$

$X_5, d_5 = 2$

$X_{13}, d_{13} = 1$ $X_{14}, d_{14} = 2$ $X_{11}, d_{11} = 1$

$X_3, d_3 = 2$

$X_7, d_7 = 1$ $X_4, d_4 = 1$

$X_9, d_9 = 2$

$X_{10}, d_{10} = 2$ $X_8, d_8 = 2$ $X_6, d_6 = 1$

Multivariate species sampling processes

Given the sampling mechanism of the Bayesian model

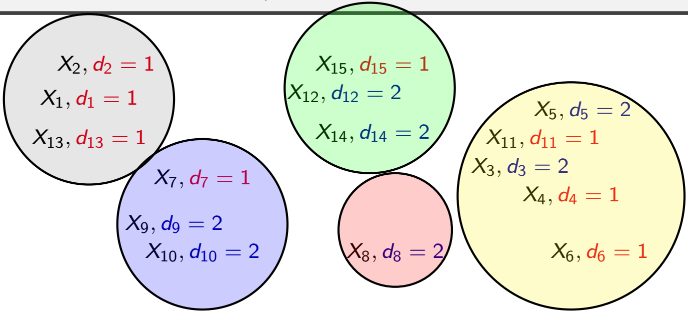
$$X_i \mid (\tilde{P}_1, \ldots, \tilde{P}_J) \overset{ind}{\sim} \tilde{P}_{d_i} \qquad (\tilde{P}_1, \ldots, \tilde{P}_J) \sim Q \tag{1}$$

Definition mSSP

$(\tilde{P}_1, \ldots, \tilde{P}_J)$ is a **multivariate species sampling process (mSSP)**,
if sampling $\boldsymbol{X}$ from (1) is equivalent to

step. 1 sampling a **partially exchangeable random partition** Π_n,

step. 2 independently sampling unique values $\{X_1^*, \ldots, X_K^*\}$ for each set from a non-atomic prob. measure P_0.



Multivariate species sampling processes

Given the sampling mechanism of the Bayesian model

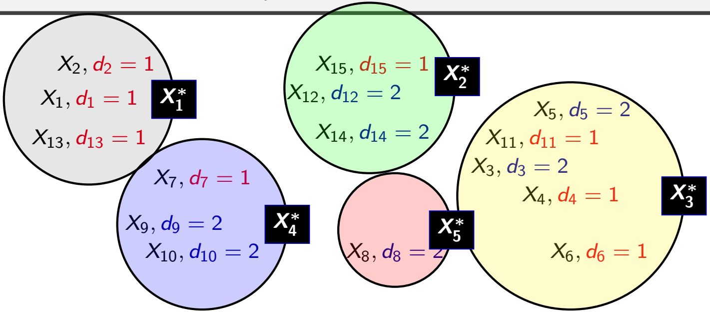
$$X_i \mid (\tilde{P}_1, \dots, \tilde{P}_J) \overset{ind}{\sim} \tilde{P}_{d_i} \qquad (\tilde{P}_1, \dots, \tilde{P}_J) \sim Q \tag{1}$$

Definition mSSP

$(\tilde{P}_1, \dots, \tilde{P}_J)$ is a **multivariate species sampling process (mSSP)**, if sampling $\boldsymbol{X}$ from (1) is equivalent to

step. 1 sampling a **partially exchangeable random partition** Π_n,

step. 2 independently sampling unique values $\{X_1^*, \dots, X_K^*\}$ for each set from a non-atomic prob. measure P_0.



Multivariate species sampling processes

Characterization theorem 1

Given $\mathbf{d}$, Π is a partially exchangeable random partition if and only if it exists a function f s.t.

$$\mathbb{P}(\Pi_n = \{A_1, \dots, A_K\}) = f\begin{pmatrix} n_{1,1} & n_{1,2} & \dots & n_{1,j} & \dots & n_{1,J} \\ n_{2,1} & n_{2,2} & \dots & n_{2,j} & \dots & n_{2,J} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ n_{k,1} & n_{k,2} & \dots & n_{k,j} & \dots & n_{k,J} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ n_{K,1} & n_{K,2} & \dots & n_{K,j} & \dots & n_{K,J} \end{pmatrix}$$

where f is a function satisfying the three following conditions:

(f-i) $f : \bigcup_{n=1}^{+\infty} \bar{\rho}_n^*(I_1, \dots, I_J) \to [0, 1]$, ($\Rightarrow$ sufficiency of the matrix of counts)

(f-ii) $f(1) = 1$ and $f(\boldsymbol{n}) = \sum_{l=1}^{K+1} f(\boldsymbol{n}^{lj+})$, ($\Rightarrow$ close to marginalization)

(f-iii) $f((\boldsymbol{n}_1, \dots, \boldsymbol{n}_J)) = f((\alpha(\boldsymbol{n}_1), \dots, \alpha(\boldsymbol{n}_J)))$ ($\Rightarrow$ invariance to sets labels)

f is called **partially exchangeable partition probability function** (pEPPF).

Characterization theorem 2

$(\tilde{P}_1, \ldots, \tilde{P}_J)$ is a mSSP if and only if

$$\tilde{P}_j \overset{a.s.}{=} \sum_{h \geq 1} \pi_{j,h} \delta_{\theta_h} + \left(1 - \sum_{h \geq 1} \pi_{j,h}\right) P_0, \qquad \text{for } j = 1, \ldots, J,$$

where θ_h are i.i.d from P_0 and independent from $\boldsymbol{\pi} = (\pi_{j,h})_{j,h}$.

Characterization theorem 2

$(\tilde{P}_1, \dots, \tilde{P}_J)$ is a mSSP if and only if

$$\tilde{P}_j \overset{a.s.}{=} \sum_{h \geq 1} \pi_{j,h} \delta_{\theta_h} + \left(1 - \sum_{h \geq 1} \pi_{j,h} \right) P_0, \qquad \text{for } j = 1, \dots, J,$$

where θ_h are i.i.d from P_0 and independent from $\boldsymbol{\pi} = (\pi_{j,h})_{j,h}$.

Characterization theorem 3

$(\tilde{P}_1, \dots, \tilde{P}_J)$ is a mSSP if and only if the predictive distribution of $\boldsymbol{X}$ is given by:

$$X_1 \sim P_0 \qquad X_{n+1} \mid \boldsymbol{X}_{1:n} \sim \sum_{l=1}^{K} p_{d_{n+1},l}(\boldsymbol{n})\, \delta_{X_l^*} + p_{d_{n+1},K+1}(\boldsymbol{n})\, P_0 \tag{2}$$

where

(p-i) $p_{j,l}(\boldsymbol{n}) \geq 0$,

(p-ii) $\sum_{l=1}^{K+1} p_{j,l}(\boldsymbol{n}) = 1$, $\forall\, \boldsymbol{n}$ and $\forall j = 1, \dots J$,

(p-iii) $p_{j,l}(\boldsymbol{n}) p_{j',r}(\boldsymbol{n}^{lj+}) = p_{j',r}(\boldsymbol{n}) p_{j,l}(\boldsymbol{n}^{rj'+})$, $\forall j, j' \in \{1, \dots, J\}$ and $\forall l, r$,

(p-iv) $p_{j,l}((\boldsymbol{n}_1, \dots, \boldsymbol{n}_J)) = p_{j,\alpha^{-1}(l)}((\alpha(\boldsymbol{n}_1), \dots, \alpha(\boldsymbol{n}_J)))$.

The collection of functions $p_{j,l}$ is called **multivariate prediction probability function** (mPPF).

Correlation structure implied by mSSM

Warning: Change of notation for observations for sake of clarity.

Correlation structure implied by a mSSM

If $(\tilde{P}_1, \tilde{P}_2) \sim$ mSSP, then, for any A s.t. $0 < P_0(A) < 1$, we obtain

$$\mathrm{corr}\{\tilde{P}_1(A), \tilde{P}_2(A)\} = \frac{\mathbb{P}(X_{1,1} = X_{2,1})}{\sqrt{\mathbb{P}(X_{1,1} = X_{1,2})}\sqrt{\mathbb{P}(X_{2,1} = X_{2,2})}} \geq 0.$$

Moreover, if the marginal distributions of $(\tilde{P}_1, \tilde{P}_2)$ are equal, we get

$$\mathrm{corr}\{\tilde{P}_1(A), \tilde{P}_2(A)\} = \frac{\mathbb{P}(X_{1,1} = X_{2,1})}{\mathbb{P}(X_{1,1} = X_{1,2})} = \frac{\mathbb{P}(\text{“ tie across groups ”})}{\mathbb{P}(\text{“ tie within a group ”})}$$

Take home messages:

- Rather than saying that “typically” $\mathrm{corr}\{\tilde{P}_1(A), \tilde{P}_2(A)\}$ does not depend on A, now we know it is true for the whole class of mSSP $\Longrightarrow$ it is a function of probabilities of sharing atoms regardless of their specific values
- $\mathrm{corr}\{\tilde{P}_1(A), \tilde{P}_2(A)) \geq 0$ and also

$$\mathrm{corr}(X_{1,1}, X_{2,1}) = \mathbb{P}(X_{1,1} = X_{2,1}) \geq 0.$$

- The dependence boils down to the sharing of underlying common atoms.

Table: Correlation, tie probabilities and extreme cases.

Process	Correlation	$P(\text{Ties Across})$	$P(\text{Ties Within})$	Indep.	Exchang.
HDP	$\frac{1+\alpha}{1+\alpha+\alpha_0}$	$\frac{1}{1+\alpha_0}$	$\frac{1+\alpha+\alpha_0}{(1+\alpha)(1+\alpha_0)}$	$\alpha_0 \to +\infty$	$\alpha \to +\infty$
HPY	$\frac{(1+\alpha)(1-\sigma_0)}{(1-\sigma\sigma_0)+\alpha(1-\sigma_0)+\alpha_0(1-\sigma)}$	$\frac{1-\sigma_0}{1+\alpha_0}$	$\frac{(1-\sigma\sigma_0)+\alpha(1-\sigma_0)+\alpha_0(1-\sigma)}{(1+\alpha)(1+\alpha_0)}$	$\alpha_0 \to +\infty$ or $\sigma_0 \to 1$	$\alpha \to +\infty$ or $\sigma \to 1$
HDM	$\frac{(1+\tau_0)(1+\tau M)}{(1+\tau M)(1+\tau_0 M_0)-\tau\tau_0(M-1)(M_0-1)}$	$\frac{1+\tau_0}{1+\tau_0 M_0}$	$\frac{(1+\tau M)(1+\tau_0 M_0)-\tau\tau_0(M-1)(M_0-1)}{(1+\tau M)(1+\tau_0 M_0)}$	$M_0 \to +\infty$	$M \to +\infty$
HGN	$\frac{\gamma_0(\gamma+1)}{(\gamma+\gamma_0)}$	$\frac{2\gamma_0}{\gamma_0+1}$	$\frac{2(\gamma+\gamma_0)}{(\gamma+1)(\gamma_0+1)}$	$\gamma_0 \to 0$	$\gamma \to 0$
HSSP	$\frac{\mathrm{EPPF}^{(2)}_{1,0}(2)}{\mathrm{EPPF}^{(2)}_{1,1}(2)+\mathrm{EPPF}^{(2)}_{2,1}(1,1)\mathrm{EPPF}^{(2)}_{1,0}(2)}$ *	$\mathrm{EPPF}^{(2)}_{1,0}(2)$	$\mathrm{EPPF}^{(2)}_{1,1}(2)+\mathrm{EPPF}^{(2)}_{2,1}(1,1)\mathrm{EPPF}^{(2)}_{1,0}(2)$	$\mathrm{EPPF}^{(2)}_{1,0}(2)=0$	$\mathrm{EPPF}^{(2)}_{1,1}(2)=0$
NDP	$\frac{1}{1+\alpha}$	$\frac{1}{(1+\alpha)(1+\beta)}$	$\frac{1}{1+\beta}$	$\alpha \to +\infty$	$\alpha \to 0$
NPY	$\frac{1-\sigma_\alpha}{1+\alpha}$	$\frac{(1-\sigma_\alpha)(1-\sigma_\beta)}{(1+\alpha)(1+\beta)}$	$\frac{1-\sigma_\beta}{1+\beta}$	$\alpha \to +\infty$ or $\sigma_\alpha \to 1$	$(\alpha,\sigma_\alpha) \to$ $\to (0,0)$
NDM	$\frac{1+\tau_\alpha}{1+\tau_\alpha M_\alpha}$	$\frac{(1+\tau_\alpha)(1+\tau_\beta)}{(1+\tau_\alpha M_\alpha)(1+\tau_\beta M_\beta)}$	$\frac{1+\tau_\beta}{1+\tau_\beta M_\beta}$	$M_\alpha \to +\infty$	$M_\alpha \to 1$
NGN	$\frac{2\gamma_\alpha}{\gamma_\alpha+1}$	$\frac{4\gamma_\alpha\gamma_\beta}{(\gamma_\alpha+1)(\gamma_\beta+1)}$	$\frac{2\gamma_\beta}{\gamma_\beta+1}$	$\gamma_\alpha \to 0$	$\gamma_\alpha \to 1$
NSSP	$\mathrm{EPPF}^{(2)}_{1,0}(2)$	$\mathrm{EPPF}^{(2)}_{1,0}(2)\mathrm{EPPF}^{(2)}_{1,1}(2)$	$\mathrm{EPPF}^{(2)}_{1,1}(2)$	$\mathrm{EPPF}^{(2)}_{1,0}(2)=0$	$\mathrm{EPPF}^{(2)}_{1,0}(2)=1$
+DP	$\frac{\frac{\epsilon_j\epsilon_k}{1+\alpha_0}}{\sqrt{\left(\frac{\epsilon_j^2}{1+\alpha_0}+\frac{(1-\epsilon_j)^2}{1+\alpha_j}\right)\left(\frac{\epsilon_k^2}{1+\alpha_0}+\frac{(1-\epsilon_k)^2}{1+\alpha_k}\right)}}$	$\frac{\epsilon_j\epsilon_k}{1+\alpha_0}$	$\frac{\epsilon_j^2}{1+\alpha_0}+\frac{(1-\epsilon_j)^2}{1+\alpha_j}$	$\epsilon=0$	$\epsilon=1$
+PY	$\frac{\frac{\epsilon_j\epsilon_k(1-\sigma_0)}{1+\alpha_0}}{\sqrt{\left(\frac{\epsilon_j^2(1-\sigma_0)}{1+\alpha_0}+\frac{(1-\epsilon_j)^2(1-\sigma_j)}{1+\alpha_j}\right)\left(\frac{\epsilon_k^2(1-\sigma_0)}{1+\alpha_0}+\frac{(1-\epsilon_k)^2(1-\sigma_k)}{1+\alpha_k}\right)}}$	$\frac{\epsilon_j\epsilon_k(1-\sigma_0)}{1+\alpha_0}$	$\frac{\epsilon_j^2(1-\sigma_0)}{1+\alpha_0}+\frac{(1-\epsilon_j)^2(1-\sigma_j)}{1+\alpha_j}$	$\epsilon=0$	$\epsilon=1$
+DM	$\frac{\frac{\epsilon_j\epsilon_k(1+\tau_0)}{1+\tau_0 M_0}}{\sqrt{\left(\frac{\epsilon_j^2(1+\tau_0)}{1+\tau_0 M_0}+\frac{(1-\epsilon_j)^2(1+\tau_j)}{1+\tau_j M_j}\right)\left(\frac{\epsilon_k^2(1+\tau_0)}{1+\tau_0 M_0}+\frac{(1-\epsilon_k)^2(1+\tau_k)}{1+\tau_k M_k}\right)}}$	$\frac{\epsilon_j\epsilon_k(1+\tau_0)}{1+\tau_0 M_0}$	$\frac{\epsilon_j^2(1+\tau_0)}{1+\tau_0 M_0}+\frac{(1-\epsilon_j)^2(1+\tau_j)}{1+\tau_j M_j}$	$\epsilon=0$	$\epsilon=1$
+GN	$\frac{\frac{\epsilon_j\epsilon_k 2\gamma_0}{\gamma_0+1}}{\sqrt{\left(\frac{\epsilon_j^2 2\gamma_0}{\gamma_0+1}+\frac{(1-\epsilon_j)^2 2\gamma_j}{\gamma_j+1}\right)\left(\frac{\epsilon_k^2 2\gamma_0}{\gamma_0+1}+\frac{(1-\epsilon_k)^2 2\gamma_k}{\gamma_k+1}\right)}}$	$\frac{\epsilon_j\epsilon_k 2\gamma_0}{\gamma_0+1}$	$\frac{\epsilon_j^2 2\gamma_0}{\gamma_0+1}+\frac{(1-\epsilon_j)^2 2\gamma_j}{\gamma_j+1}$	$\epsilon=0$	$\epsilon=1$
+SSP	$\frac{\epsilon_j\epsilon_k\mathrm{EPPF}^{(2)}_{1,0}(2)}{\sqrt{(\epsilon_j^2\mathrm{EPPF}^{(2)}_{1,0}(2)+(1-\epsilon_j)^2\mathrm{EPPF}^{(2)}_{1,1}(2))(\epsilon_k^2\mathrm{EPPF}^{(2)}_{1,0}(2)+(1-\epsilon_k)^2\mathrm{EPPF}^{(2)}_{1,1}(2))}}$	$\epsilon_j\epsilon_k\mathrm{EPPF}^{(2)}_{1,0}(2)$	$\epsilon_j^2\mathrm{EPPF}^{(2)}_{1,0}(2)+(1-\epsilon_j)^2\mathrm{EPPF}^{(2)}_{1,1}(2)$	$\epsilon=0$	$\epsilon=1$
GM-DP	$\frac{(1-z)c}{1+c}\,{}_3F_2(a,1,1;b,b;1)^{\star}$	$\frac{(1-z)c}{(1+c)^2}\,{}_3F_2(a,1,1;b,b;1)^{\star\star}$	$\frac{1}{1+c}$	$z=1$	$z=0$
GM-σ	$(1-z)\mathcal{I}(c,z)^{\star\star\star}$	$(1-z)(1-\sigma)\mathcal{I}(c,z)$	$1-\sigma$		
HHDP	$1-\frac{\alpha\beta_0}{(1+\alpha)(\beta_0+\beta+1)}$	$\frac{1}{\beta_0+1}+\frac{\beta_0}{(1+\alpha)(1+\beta)(1+\beta_0)}$	$\frac{1+\beta+\beta_0}{(1+\beta)(1+\beta_0)}$	$(\alpha,\beta_0) \to$ $\to (+\infty,+\infty)$	$\alpha \to 0$
nCAM	$1-\frac{\beta\alpha}{(2\beta+1)(1+\alpha)}$	$\frac{1}{1+\alpha}\left(\frac{1}{1+\beta}+\frac{\alpha}{2\beta+1}\right)$	$\frac{1}{1+\beta}$	None	$\alpha \to 0$

* $\mathrm{EPPF}_{\cdot,1}$ and $\mathrm{EPPF}_{\cdot,0}$ are induced by $\mathcal{L}_{\pi,1}=\ldots=\mathcal{L}_{\pi,J}$ and $\mathcal{L}_{\pi,0}$, respectively. ** ${}_3F_2$ is the generalized hypergeometric function and $a=\alpha(1-z)+2$, $b=\alpha+2$ *** $\mathcal{I}(c,z)=\frac{1}{\sigma}\int_0^1\frac{w^{1/\sigma-1}}{[1+z(1-\omega^{1/\sigma})^{\sigma}-z(1-\omega)]}d\omega$

Correlation structure

Questions: Are there models for which:

- $\mathrm{corr}(\tilde{P}_1(A), \tilde{P}_2(A)) = 1$ if and only if $\tilde{P}_1 = \tilde{P}_2$ a.s. i.e. full exchangeability?
- $\mathrm{corr}(\tilde{P}_1(A), \tilde{P}_2(A)) = 0$ if and only if $\tilde{P}_1 \perp\!\!\!\perp \tilde{P}_2$?

Extreme cases

Under mild conditions which are satisfied by all dependent processes used in Bayesian statistics:

- $\mathrm{corr}(\tilde{P}_1(A), \tilde{P}_2(A)) = 1$ if and only if $\tilde{P}_1 = \tilde{P}_2$ a.s.
  i.e. full exchangeability;
- $\mathrm{corr}(\tilde{P}_1(A), \tilde{P}_2(A)) = 0$ if and only if $\tilde{P}_1 \perp\!\!\!\perp \tilde{P}_2$.

$\Longrightarrow$ justifies the use of correlation as measure of dependence.

Predictive schemes

If $\boldsymbol{X} \sim mSSM$ with pEPPF, then a multivariate Chinese restaurant process (mCRP) (i.e., a sequential sampling scheme that allows sampling from the predictive distribution) can be derived as

$$X_{j,I_j+1} \mid (\boldsymbol{X}_{j,1:I_j})_{j=1}^J = \begin{cases} X_l^* & \text{w.p. } \dfrac{\mathrm{pEPPF}_D^{(n+1)}(\boldsymbol{n}_1,\ldots,[n_{j,1},\ldots,n_{j,l}+1,\ldots,n_{1,D}],\ldots,\boldsymbol{n}_J)}{\mathrm{pEPPF}_D^{(n)}(\boldsymbol{n}_1,\ldots,[n_{j,1},\ldots,n_{j,l},\ldots,n_{j,D}],\ldots,\boldsymbol{n}_J)} \\ X_{new}^* & \text{w.p. } \dfrac{\mathrm{pEPPF}_{D+1}^{(n+1)}([\boldsymbol{n}_1,0],[n_{j,1},\ldots,n_{j,l},\ldots,n_{j,D},1],\ldots,[\boldsymbol{n}_J,0])}{\mathrm{pEPPF}_D^{(n)}(\boldsymbol{n}_1,\ldots,[n_{j,1},\ldots,n_{j,l},\ldots,n_{j,D}],\ldots,\boldsymbol{n}_J)} \end{cases},$$

where $(X_1^*, \ldots, X_D^*)$ are the unique values in $(\boldsymbol{X}_{j,1:I_j})_{j=1}^J$ recorded in order of arrival by group, $n = \sum_j I_j$, and X_{new}^* is a new species.

Remark: Intractable in general, but exploiting variable augmentations of pEPPF as a mixture of EPPFs we recover tractable composition of CRP (as CRF for HDP).

Augmented pEPPF

- If $(P_1, \ldots, P_J)$ is an hierarchical SSP, then

$$\text{pEPPF}^{(n)}_{D,\text{aug}}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_J, \boldsymbol{\ell}, \boldsymbol{q}) = \text{EPPF}^{(\ell_{\cdot,\cdot})}_{D,0}(\ell_{\cdot,1}, \ldots, \ell_{\cdot,D}) \prod_{j=1}^{J} \text{EPPF}^{(I_j)}_{\ell_{j,\cdot},j}(q_{j,1}, \ldots, q_{j,\ell_{j,\cdot}})$$

- If $(P_1, \ldots, P_J)$ is an nested SSP, then

$$\text{pEPPF}^{(n)}_{D,\text{aug}}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_J, \boldsymbol{\ell}, \boldsymbol{q}) = \text{EPPF}^{(J)}_{R,0}(\ell_1, \ldots, \ell_R) \prod_{r=1}^{R} \text{EPPF}^{(I^\star_r)}_{D_r}(q_{1,\cdot}, \ldots, q_{D_r,})$$

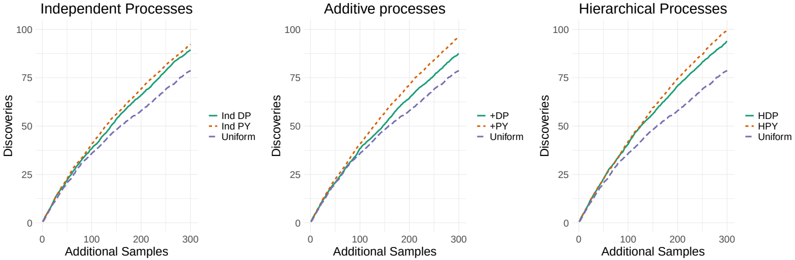
- If $(P_1, \ldots, P_J)$ is an additive SSP, then

$$\text{pEPPF}^{(n)}_{D,\text{aug}}(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_J, \boldsymbol{\ell}, \boldsymbol{q}) = \prod_{j=1}^{J} \epsilon_j^{\ell_0}(1-\epsilon_j)^{\ell_j} \prod_{j=0}^{J} \text{EPPF}^{(\ell_j)}_{D_j,j}(q_{j,1}, \ldots, q_{j,D_j})$$

- ...

rmSSP illustration

- We illustrate some rmSSP performance in devising a strategy for sequentially selecting sampling sites across various locations to maximize the diversity of observed species in a trees dataset (Battiston, Favaro & Teh, 2018).
- **Data**: species of trees observed in South America recorded from 4 groups, according to spatial location.
- The goal of maximizing the number of species discovered via sequential sampling can be formulated as a multi-armed bandit problem, where each arm is constituted by a certain site/population, and a unitary reward is gained when a new species is observed.



Number of species discovered as a function of additional sample sizes in the rmSSPs and the uniform model. The uniform model selects arms randomly.

Joint work with

Some References

Some References

- Antoniak (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2**, 1152–1174.
- Argiento, Cremaschi & Vannucci, (2020). Hierarchical normalized completely random measures to cluster grouped data. *J. Am. Stat. Assoc.*, **115**, 318–333.
- Ascolani, Lijoi, Rebaudo & Zanella (2023). Clustering consistency with Dirichlet process mixtures. *Biometrika*, **110**, 551-558.
- Balocchi, Favaro & Naulet (2024). Bayesian nonparametric inference for "species-sampling" problems. *arXiv:2203.06076*.
- Bassetti, Casarin & Rossini (2020). Hierarchical species sampling models. *Bayesian Anal.*, **15**, 809–838.
- Battiston, Favaro & Teh (2018). Multi-armed bandit for species discovery: a Bayesian nonparametric approach. *J. Am. Stat. Assoc.*, **113**, 455–466.
- Beraha, Guglielmi & Quintana (2021). The semi-hierarchical Dirichlet process and its application to clustering homogeneous distributions. *Bayesian Anal.*, **16**, 1187–1219.
- Camerlenghi, Dunson, Lijoi, Prünster & Rodriguez (2019). Latent nested nonparametric priors. (With discussion) *Bayesian Anal.* **15**, 1303-1356.
- Camerlenghi, Lijoi, Orbanz & Prünster (2019). Distribution theory for Hierarchical Processes. *Annal. Statist.* **47**, 67-92.
- Catalano, Lijoi & Prünster (2021). Measuring dependence in the Wasserstein distance for Bayesian nonparametric models. *Ann. Statist* **49** 2916-2947.
- Cifarelli & Regazzini (1978). Nonparametric statistical problems under partial exchangeability. *Quad. Ist. Matem. Fin. Univ. di Torino Ser. III* **12**, 1-36.
- De Blasi, Favaro, Lijoi, Mena, Prünster & Ruggiero (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE TPAMI* **37** 212-229.
- Denti, Camerlenghi, Guindani & Mira (2023). A common atom model for the Bayesian nonparametric analysis of nested data. *J. Am. Stat. Assoc.*, **118**, 405–416.
- Ferguson (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209-30.
- de Finetti (1931). Probabilismo. *Logos* **14**, 163-219. [translated in *Erkenntnis* **31**, 169–223, 1989].
- de Finetti (1938). Sur la condition d'équivalence partielle. *Act.sci. ind.* **739**, 5-18.

• Franzolini, B., Lijoi, A., Prünster, I., & Rebaudo, G. (2025). Multivariate species sampling models. arXiv preprint arXiv:2503.24004.

• Franzolini, B., Lijoi, A., Prünster, I., & Rebaudo, G. (2025+). Partially exchangeable random partition structures (Ongoing).

• Horiguchi, Chan & Ma (2022). Tree stick-breaking priors for covariate-dependent mixture models. *arXiv:2208.02806*.

• Lee, Quintana, Müller & Trippa (2013). Defining predictive probability functions for species sampling models. *Stat. Sci.*, **28**, 209–222

• Lijoi, Nipoti & Prünster (2014). Bayesian inference with dependent normalized completely random measures. *Bernoulli* **20**, 1260-1291.

• Lijoi, Prünster & Rebaudo (2023). Flexible clustering via hidden hierarchical Dirichlet priors. *Scand. J. Stat.*, **50**, 213-234.

• Griffin & Leisen (2017). Compound random measures and their use in Bayesian non-parametrics. *J. R. Stat. Soc. Series B Stat. Methodol.*, **79**, 525–545.

• MacEachern (1999). Dependent nonparametric processes. In *ASA Proceedings*.

• MacEachern (2000). Dependent Dirichlet processes. *OSU Tech. Report*.

• Müller, Quintana & Rosner (2004). A method for combining inference across related nonparametric Bayesian models. *J. Roy. Statist. Soc. Ser. B*, **66**, 735-749.

• Pitman (1996). Some developments of the Blackwell-MacQueen urn scheme. *IMS Lecture Notes Monogr.* **30**, 245-267.

• Pitman (2006). *Combinatorial Stochastic Processes*. Lecture Notes in Math., vol.1875, Springer, Berlin.

• Quintana, Müller, Jara & MacEachern (2022). The dependent Dirichlet process and related models. *Stat. Sci.*, **37**, 24–41.

• Regazzini (1978). Intorno ad alcune questioni relative alla definizione del premio secondo la teoria della credibilità. *Giorn. Istit. Ital. Attuari*, **41**, 77–89.

• Rodríguez, Dunson & Gelfand (2008). The nested Dirichlet process. *J. Amer. Statist. Assoc.* **103**, 1131-1144.

• Teh & Jordan (2010). Hierarchical Bayesian nonparametric models with applications. In *Bayesian Nonparametrics*, 158–207. Cambridge Univ. Press.

• Teh, Jordan, Beal & Blei (2006). Hierarchical Dirichlet processes. *J. Amer. Statist. Assoc.* **101**, 1566–1581.

Notation

$\rho_K(l_1,\ldots,l_J) = \{\boldsymbol{n} : \boldsymbol{n} \in \mathbb{N}_0^{K\times J},\ \sum_{l=1}^{K} n_{l,j} = l_j \text{ and } \sum_{j=1}^{J} n_{l,j} > 0 \text{ for } l = 1,\ldots,K \text{ and } j = 1,\ldots,J\}$

However, not all matrices in $\rho_K(l_1,\ldots,l_J)$ correspond to a partition (in order of appearance), when such partition exists we say that $\boldsymbol{n}$ is a *compatible matrix of counts* accordingly to $\mathscr{D}_n$ (or, shortly, $\boldsymbol{n}$ is $\mathscr{D}_n$-compatible). To clarify why not all the matrices in $\rho_K(l_1,\ldots,l_J)$ correspond to a partition, we provide the following two examples.

Example Let $n = 7$, $d = \{1,1,1,2,2,2,2\}\}$ and K=2. The matrix

$$\boldsymbol{n} = \begin{pmatrix} 0 & 3 \\ 3 & 1 \end{pmatrix}$$

is not a compatible matrix of counts accordingly to d. The order of appearance requires $n_{1,1} > 0$.

Example Let $n = 7$, $D_n = (1,2,1,1,2,2,2)$ and K=3. The matrix

$$\boldsymbol{n} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 4 \end{pmatrix}$$

is not a compatible matrix of counts accordingly to $\mathscr{D}_n$. The order of appearance requires $n_{1,2} + n_{2,2} > 0$.

Finally, we denote with

$$\rho_K^*(l_1,\ldots,l_J) = \{\boldsymbol{n} : \boldsymbol{n} \in \rho_K(l_1,\ldots,l_J),\ \boldsymbol{n} \text{ is } \mathscr{D}_n\text{-compatible}\}$$

and

$$\bar{\rho}_n^*(l_1,\ldots,l_J) = \bigcup_{K=1}^{n} \rho_K^*(q_1,\ldots,q_J).$$