The Stirling-gamma process and its application to Bayesian nonparametrics

Tommaso Rigon

Joint work with: Alessandro Zito and David B. Dunson

13th Bayesian inference for Stochastic Processes

(Real Academia de Ciencias, Madrid)

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

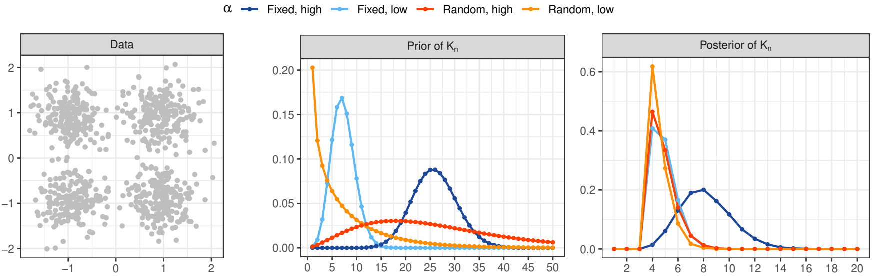
Introduction

- Discrete Bayesian nonparametric **priors** are widely used tools for clustering, density estimation, and species discovery.
- Notable examples are the **Dirichlet process** (DP) and the **Pitman–Yor** (PY).
- It is common to consider a hierarchical specification of the kind

$$(\tilde{p} \mid \alpha) \sim \text{DP}(\alpha P), \qquad \alpha \sim \pi(\alpha),$$

  to learn the **precision parameter** of the Dirichlet process.
- This is particularly relevant for **mixture models**, as it increases the **robustness** of the prior specification.
- In a seminal JASA paper, Escobar and West (1995) used $\alpha \sim \text{Ga}(a, b)$.
- This talk is about an **interpretable** and (sometimes) **conjugate** prior for α.

A robustness issue



- A common Bayesian nonparametric **mixture model** is

$$X_i \mid \theta_i \overset{\text{ind}}{\sim} f(x \mid \theta_i), \qquad \theta_i \mid \tilde{p} \overset{\text{iid}}{\sim} \tilde{p}, \qquad \tilde{p} \sim \mathscr{Q}, \qquad (i = 1, \dots, n),$$

where $\theta_1, \dots, \theta_n$ are latent parameters.

- Center/right panel: prior/posterior distribution of the **number of clusters** under a **Dirichlet** and a **Stirling gamma** process.

Discrete random structures

- Let us consider a set of exchangeable random variables $\theta_1, \dots, \theta_n$, namely

$$(\theta_i \mid \tilde{p}) \overset{\text{iid}}{\sim} \tilde{p}, \qquad i = 1, \dots, n,$$
$$\tilde{p} \sim \mathbf{Q}.$$

- The probability measure $\mathbf{Q}$ represents the prior law.

- A species sampling model is a **discrete random probability measure**, so that

$$\tilde{p} = \sum_{h=1}^{\infty} \pi_h \delta_{Z_h}, \qquad Z_h \overset{\text{iid}}{\sim} P,$$

independently on the **random probabilities** $(\pi_1, \pi_2, \dots)$, with P **diffuse**.

- Well-known Gibbs-type priors are recovered: the Dirichlet process, the Pitman–Yor process, and the normalized generalized Gamma process.

Gibbs-type priors

- The discreteness of $\tilde{p}$ implies that there will be ties among observations $\theta_1, \dots, \theta_n$, therefore inducing a **random partition**, say Ψ_n.

- In **Gibbs-type priors** a specific partition of the integers $\{1, \dots, n\}$ into k sets $C_1, \dots, C_k$ is regulated by the EPPF, which has a **product form**:

$$\Pi(n_1, \dots, n_k) = \mathrm{pr}(\Psi_n = \{C_1, \dots, C_k\}) = V_{n,k} \prod_{j=1}^{k} (1 - \sigma)_{n_j - 1},$$

  with $\sigma < 1$, $n_j = \mathrm{card}(C_j)$ and $\sum_{j=1}^{k} n_j = n$.

- The non-negative weights $V_{n,k}$ satisfy the forward recursive equation

$$V_{n,k} = (n - \sigma) V_{n+1,k} + V_{n+1,k+1},$$

  for any $k = 1, \dots, n$ and $n \geq 1$, with $V_{1,1} = 1$.

Gibbs-type priors

- The **predictive distribution** of θ_{n+1}, conditional on $\theta^{(n)} = (\theta_1, \dots, \theta_n)$ has a simple form:

$$\mathbb{P}(\theta_{n+1} \in A \mid \theta^{(n)}) = \frac{V_{n+1,k+1}}{V_{n,k}} P(A) + \frac{V_{n+1,k}}{V_{n,k}} \sum_{j=1}^{k} (n_j - \sigma) \delta_{\theta_j^*}(A).$$

- Moreover, the number K_n of **distinct values** in $\theta^{(n)}$ has probability distribution

$$\mathbb{P}(K_n = k) = V_{n,k} \frac{\mathscr{C}(n, k; \sigma)}{\sigma^k},$$

with $\mathscr{C}(n, k; \sigma)$ denoting the generalized factorial coefficient.

- The random variable K_n is of great interest e.g. in mixture models, as it denotes the **number of clusters** we expect **a priori**.

The $\sigma = 0$ case

- The **Dirichlet process** is an instance of Gibbs-type prior with $\sigma = 0$. Indeed:
- The EPPF of the Dirichlet process is

$$\Pi(n_1, \dots, n_k \mid \alpha) = \frac{\alpha^k}{(\alpha)_n} \prod_{j=1}^{k} (n_j - 1)!.$$

- The **urn-scheme** (Blackwell and MacQueen, 1973) is

$$\mathbb{P}(\theta_{n+1} \in A \mid \theta^{(n)}) = \frac{\alpha}{\alpha + n} P(A) + \frac{1}{\alpha + n} \sum_{j=1}^{k} n_j \delta_{\theta_j^*}(A).$$

- The distribution of the **number of clusters** (Antoniak 1974) is

$$\mathbb{P}(K_n = k \mid \alpha) = \frac{\alpha^k}{(\alpha)_n} |s(n,k)|, \qquad \mathbb{E}(K_n \mid \alpha) = \sum_{i=1}^{n} \frac{\alpha}{\alpha + i - 1},$$

with $s(n, k)$ denoting the Stirling number of the first kind.

The $\sigma = 0$ case

- As shown in the example, the distribution of K_n is **highly concentrated**.
- Therefore, in order to **robustify** inference, one could place a prior on α.
- Placing a prior on α has a **remarkable connection** with Gibbs-type priors with $\sigma = 0$.
- The $V_{n,k}$ of a Gibbs-type priors with $\sigma = 0$ can be **always represented** as

$$V_{n,k} = \int_{\mathbb{R}^+} \frac{\alpha^k}{(\alpha)_n} \pi(\alpha) \mathrm{d}\alpha,$$

  for some probability distribution $\pi(\alpha)$, a result due to Gnedin and Pitman (2005).
- What it is a **natural candidate** for $\pi(\alpha)$? Under a Gamma prior, the resulting marginal **properties** are **unclear**...

The Stirling-gamma prior

- We propose to use the **Stirling-gamma** prior, denoted $\alpha \sim \text{Sg}(a, b, m)$

$$\pi(\alpha) = \frac{1}{\mathcal{S}_{a,b,m}} \frac{\alpha^{a-1}}{\{(\alpha)_m\}^b}, \qquad \mathcal{S}_{a,b,m} = \int_{\mathbb{R}^+} \frac{\alpha^{a-1}}{\{(\alpha)_m\}^b} \mathrm{d}\alpha.$$

  where the hyperparameters $a, b > 0$ and $m \in \mathbb{N}$ satisfy the constraints $1 < a/b < m$.

- **Proposition**. The above density function is proper ($\mathcal{S}_{a,b,m} < \infty$). Moreover, iid samples can be easily obtained using the ratio of uniforms method.

- This prior for α leads to a **Gibbs-type prior** with weights

$$V_{n,k} = \frac{\mathscr{V}_{a,b,m}(n,k)}{\mathscr{V}_{a,b,m}(1,1)}, \qquad \mathscr{V}_{a,b,m}(n,k) = \int_{\mathbb{R}_+} \frac{\alpha^{a+k-1}}{\{(\alpha)_m\}^b (\alpha)_n} \mathrm{d}\alpha.$$

- Moreover, if $a, b \in \mathbb{N}$, then the above integral is **explicitly available**.

Parameter interpretation

Theorem (Zito et al., 2023+)

Let $\alpha \sim \mathrm{Sg}(a, b, m)$ and $\mathcal{D}_{a,b,m} = \mathbb{E}\{\sum_{i=0}^{m-1} \alpha^2/(\alpha + i)^2\}$. The number of clusters K_m obtained from $\theta_1, \dots, \theta_m$ is distributed as

$$\mathbb{P}(K_m = k) = \frac{\mathscr{V}_{a,b,m}(m, k)}{\mathscr{V}_{a,b,m}(1, 1)} |s(m, k)|,$$

*for $k = 1, \dots, m$, with **mean** and **variance** equal to*

$$\mathbb{E}(K_m) = \frac{a}{b}, \qquad \mathrm{var}(K_m) = \frac{b+1}{b}\left(\frac{a}{b} - \mathcal{D}_{a,b,m}\right).$$

- It can be shown that $\mathcal{D}_{a,b,m} \approx 1$ for m large enough.
- Hence, a/b is the **location**, b controls the **precision** and m is a **reference sample size**.

Limiting behavior

Theorem (Zito et al., 2023+)

Let $\alpha \sim \mathrm{Sg}(a, b, m)$. *Then, the following convergence in distribution holds:*

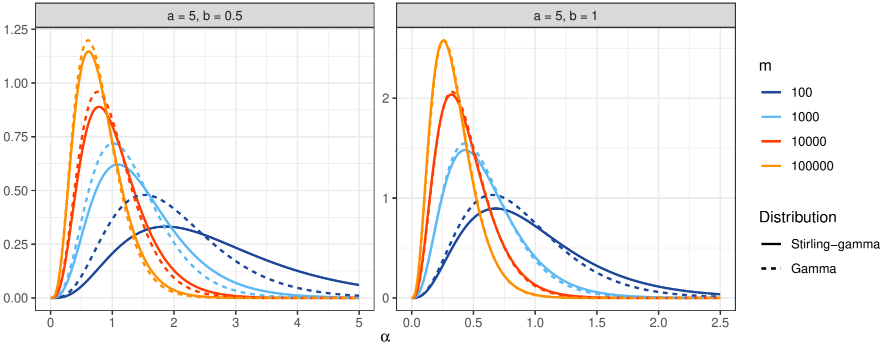
$$\alpha \log m \to \gamma, \quad \gamma \sim \mathrm{Ga}(a - b, b), \quad m \to \infty,$$

implying that $\alpha \to 0$. *Moreover, the following convergence in distribution holds:*

$$K_m \to K_\infty, \quad K_\infty \sim 1 + \mathrm{Negbin}\left(\frac{b}{b+1}, a - b\right), \quad m \to \infty.$$

- **Remark**. When m is **fixed**, it is well-known that $K_n / \log n \to \alpha \sim \mathrm{Sg}(a, b, m)$ in distribution as $n \to \infty$.
- (Very) roughly speaking, we will say that the convergence $\alpha \to 0$ counterbalances the divergence of K_n.
- In the Dirichlet process case, if $\alpha = \lambda / \log m$ for some $\lambda > 0$, then $K_m \to K_\infty$, with $K_\infty \sim 1 + \mathrm{Po}(\lambda)$ as $m \to \infty$. Thus, Stirling-gamma prior improves the **robustness**.

Graphical representation



- Density function of a $\mathrm{Sg}(a, b, m)$ (solid lines) and a $\mathrm{Ga}(a - b, b \log m)$ (dashed lines).

Exponential families: the $m = n$ case

- A simplification occurs when $m = n$, i.e. the prior depends on the sample size.
- The **key observation** is noticing that for any $n \geq 1$ the distribution

$$\mathbb{P}(K_n = k \mid \alpha) = \frac{\alpha^k}{(\alpha)_n}|s(n,k)|,$$

is an **exponential family**, with natural parameter $\psi = \log \alpha$.

- Indeed, we can equivalently write

$$\mathbb{P}(K_n = k \mid \psi) = |s(n,k)| \exp\left\{k\psi - \mathcal{K}(\psi)\right\}, \qquad \psi = \log \alpha,$$

where the **cumulant generating function** is $\mathcal{K}(\psi) = \log \Gamma(e^\psi + n) - \log \Gamma(e^\psi)$.

- **Side comment**. The properties of exponential families lead to an alternative proof of the identity

$$\mathbb{E}(K_n \mid \psi) = \sum_{i=1}^{n} \frac{e^\psi}{e^\psi + i - 1} = \frac{\partial}{\partial \psi}\mathcal{K}(\psi).$$

Diaconis and Ylvisaker priors

- **Key result**. The prior $\alpha \sim \mathrm{Sg}(a, b, n)$ is the Diaconis and Ylvisaker **conjugate prior** for the exponential family model $\mathbb{P}(K_n = k \mid \alpha)$. Note that we let $m = n$.

- A direct application of Bayes theorem leads

$$\pi(\alpha \mid K_n = k) \propto \pi(\alpha)\mathbb{P}(K_n = k \mid \alpha) \propto \frac{\alpha^{a-1}}{\{(\alpha)_n\}^b} \frac{\alpha^k}{(\alpha)_n}.$$

- Hence, the **posterior density** has the form

$$\pi(\alpha \mid K_n = k) = \frac{1}{\mathcal{S}_{a+k,b+1,n}} \frac{\alpha^{a+k-1}}{\{(\alpha)_n\}^{b+1}}.$$

- **Remark**. The number of distinct values k is the minimal **sufficient statistics** in the EPPF of the Dirichlet process.

- Hence, all the posterior findings based on the model $\mathbb{P}(K_n = k \mid \alpha)$ coincide with those based on $\Pi(n_1, \ldots, n_k \mid \alpha)$, because the likelihood contribution is the same.

The role of the hyperparameters: conjugate case

- Let $\alpha \sim \text{DY}(a, b, n)$. Then, Theorem 2 of Diaconis and Ylvisaker (1979) ensures that

$$\mathbb{E}(K_n = k) = \sum_{i=1}^{n} \mathbb{E}\left(\frac{\alpha}{\alpha + i - 1}\right) = \frac{a}{b}.$$

- Thanks to conjugacy, we can also obtain the **posterior mean** for the number of expected clusters, namely

$$\sum_{i=1}^{n} \mathbb{E}\left(\frac{\alpha}{\alpha + i - 1} \mid \theta_1, \dots, \theta_n\right) = \left(\frac{a}{b}\right) \frac{b}{b+1} + k \frac{1}{b+1}.$$

- The posterior is a convex combination of the **prior mean** a/b and the **observed number clusters** k.

- This relationship clarifies that b is a **precision** parameter, quantifying the weight of the prior with respect to the data.

Conjugacy and projectivity

- The prior dependency on the same size n has some important consequences on the process, which must be handled with care.

- The **Gibbs-type recursion** characterizing the coefficients $V_{n,k}$ **no longer holds**, namely

$$V_{n,k} \neq nV_{n+1,k} + V_{n+1,k+1}$$

- This **breaks** the **predictive scheme**, causing the sequence to lose the **projectivity** property typical of species sampling models.

- This is a limitation if the focus is on **extrapolating** $(K_{n+m} \mid K_n = k)$ from a sample to the general population, but less so on **clustering problems**.

- Indeed, several other existing priors for partitions are not projective (e.g. general product partition models, models for micro-clustering, etc.)

Communities in ant interaction networks

- We want to identify **community structures** in a colony of ant workers by modeling daily ant-to-ant interaction networks via **stochastic block models**.

- The data were collected by continuously monitoring six **colonies** of the **ant** *Camponotus fellah* through an automated tracking system, over a period of 41 days.

- Given a random partition of the nodes $\Pi_{n,s} = \{C_{1,s}, \dots, C_{k_s,s}\}$ in s, call $Z_{i,s}$ an auxiliary variable so that $Z_{i,s} = h$ if the node $i \in C_{h,s}$, for $i = 1, \dots, n$.

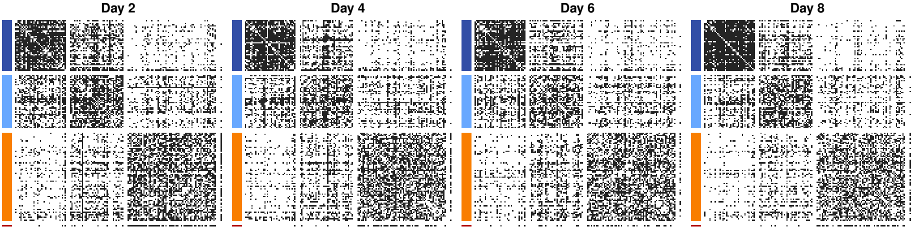
- The probability of detecting an edge between nodes i and j in network s is specified as

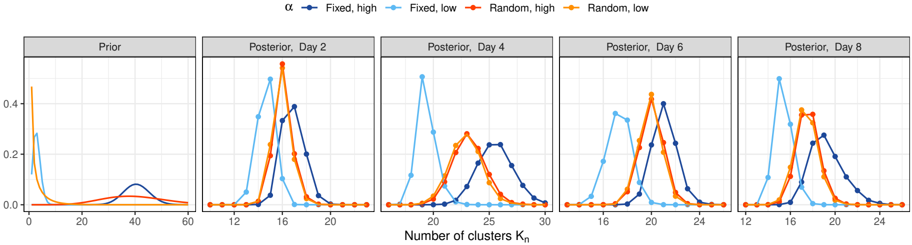
$$\mathbb{P}(X_{i,j,s} = 1 \mid Z_{i,s} = h, Z_{j,s} = h', \nu) = \nu_{h,h',s}, \quad \nu_{h,h',s} \sim \mathrm{Be}(1,1).$$

- Here, $\nu_{h,h',s}$ is the edge probability in the block identified by clusters $C_{h,s}$ and $C_{h',s}$.

- The latent partition identifies communities among the ants.

Data and results





Muchas gracias!

Escobar & West (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577–88.

De Blasi, Favaro, Lijoi, Mena, Prünster & Ruggiero (2015). Are Gibbs-Type Priors the Most Natural Generalization of the Dirichlet Process?. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37**, 212–29.

Zito, Rigon & Dunson (2023). Bayesian nonparametric modeling of latent partitions via Stirling-gamma priors. *To be submitted*