On some distributional properties of Gibbs-type priors

Igor Prünster

University of Torino & Collegio Carlo Alberto

Bayesian Nonparametrics Workshop

ICERM, 21st September 2012

Joint work with: P. De Blasi, S. Favaro, A. Lijoi and R. Mena

European Research Council
Established by the European Commission

Outline

The Bayesian nonparametric framework

de Finetti's representation theorem: a sequence of $\mathbb{X}$–valued observations $(X_n)_{n\geq 1}$ is exchangeable if and only if for any $n \geq 1$

$$\begin{aligned} X_i \mid \tilde{P} &\overset{\text{iid}}{\sim} \tilde{P} \qquad i = 1, \ldots, n \\ \tilde{P} &\sim Q \end{aligned}$$

$\Longrightarrow$ Q, defined on the space of probability measures $\mathscr{P}$, is the de Finetti measure of $(X_n)_{n\geq 1}$ and acts as a prior distribution for Bayesian inference being the law of a random probability measure $\tilde{P}$.

The Bayesian nonparametric framework

de Finetti's representation theorem: a sequence of $\mathbb{X}$–valued observations $(X_n)_{n\geq 1}$ is exchangeable if and only if for any $n \geq 1$

$$\begin{aligned} X_i \mid \tilde{P} &\overset{\text{iid}}{\sim} \tilde{P} \qquad i = 1, \dots, n \\ \tilde{P} &\sim Q \end{aligned}$$

$\Longrightarrow$ Q, defined on the space of probability measures $\mathscr{P}$, is the de Finetti measure of $(X_n)_{n\geq 1}$ and acts as a prior distribution for Bayesian inference being the law of a random probability measure $\tilde{P}$.

If Q is not degenerate on a subclass of $\mathscr{P}$ indexed by a finite dimensional parameter, it leads to a nonparametric model

$\Longrightarrow$ natural requirement (Ferguson, 1974): Q should have "large" support (possibly the whole $\mathscr{P}$)

Discrete nonparametric priors

If Q selects (a.s.) discrete distributions i.e. $\tilde{P}$ is a discrete random probability measure

$$\tilde{P}(\,\cdot\,) = \sum_{i \geq 1} \tilde{p}_i \delta_{Z_i}(\,\cdot\,), \qquad (\Diamond)$$

then a sample $(X_1, \ldots, X_n)$ will exhibit ties with positive probability i.e. feature K_n distinct observations

$$X_1^*, \ldots, X_{K_n}^*$$

with frequencies $N_1, \ldots, N_{K_n}$ such that $\sum_{i=1}^{K_n} N_i = n$.

Discrete nonparametric priors

If Q selects (a.s.) discrete distributions i.e. $\tilde{P}$ is a discrete random probability measure

$$\tilde{P}(\,\cdot\,) = \sum_{i \geq 1} \tilde{p}_i \delta_{Z_i}(\,\cdot\,), \qquad (\Diamond)$$

then a sample $(X_1, \ldots, X_n)$ will exhibit ties with positive probability i.e. feature K_n distinct observations

$$X_1^*, \ldots, X_{K_n}^*$$

with frequencies $N_1, \ldots, N_{K_n}$ such that $\sum_{i=1}^{K_n} N_i = n$.

1. Species sampling: model for species distribution within a population
   - X_i^* is the i–the distinct species in the sample;
   - N_i is the frequency of X_i^*;
   - K_n is total number of distinct species in the sample.

   $\Longrightarrow$ Species metaphor
2. Density estimation and clustering of latent variables: model for a latent level of a hierarchical model; many successful applications can be traced back to this idea due to Lo (1984) where the mixture of Dirichlet process is introduced.

Probability of discovering a new species

A key quantity is the probability of discovering a new species

$$\mathbb{P}[X_{n+1} = \text{“new”} \mid X^{(n)}] \qquad (*)$$

where throughout we set $X^{(n)} := (X_1, \ldots, X_n)$.

Probability of discovering a new species

A key quantity is the probability of discovering a new species

$$\mathbb{P}[X_{n+1} = \text{“new”} \mid X^{(n)}] \qquad (*)$$

where throughout we set $X^{(n)} := (X_1, \ldots, X_n)$.

Discrete $\tilde{P}$ can be classified in 3 categories according to $(*)$:

(a) $\mathbb{P}[X_{n+1} = \text{“new”} \mid X^{(n)}] = f(n, \text{model parameters})$

$\iff$ depends on n but not on K_n and $\mathbf{N}_n = (N_1, \ldots, N_{K_n})$

$\implies$ Dirichlet process (Ferguson, 1973);

Probability of discovering a new species

A key quantity is the probability of discovering a new species

$$\mathbb{P}[X_{n+1} = \text{“new”} \mid X^{(n)}] \qquad (*)$$

where throughout we set $X^{(n)} := (X_1, \dots, X_n)$.

Discrete $\tilde{P}$ can be classified in 3 categories according to $(*)$:

(a) $\mathbb{P}[X_{n+1} = \text{“new”} \mid X^{(n)}] = f(n, \text{model parameters})$

$\iff$ depends on n but not on K_n and $\mathbf{N}_n = (N_1, \dots, N_{K_n})$

$\implies$ Dirichlet process (Ferguson, 1973);

(b) $\mathbb{P}[X_{n+1} = \text{“new”} \mid X^{(n)}] = f(n, K_n, \text{model parameters})$

$\iff$ depends on n and K_n but not on $\mathbf{N}_n = (N_1, \dots, N_{K_n})$

$\iff$ Gibbs–type priors (Gnedin and Pitman, 2006);

(c) $\mathbb{P}[X_{n+1} = \text{“new”} \mid X^{(n)}] = f(n, K_n, \mathbf{N}_n, \text{model parameters})$

$\iff$ depends on all information conveyed by the sample i.e. n, K_n and $\mathbf{N}_n = (N_1, \dots, N_{K_n})$

$\iff$ serious tractability issues.

Complete predictive structure

$\tilde{P}$ is a Gibbs-type random probability measure of order $\sigma \in (-\infty, 1)$ if and only if it gives rise to predictive distributions of the form

$$\mathbb{P}\left[X_{n+1} \in A \,\middle|\, X^{(n)}\right] = \frac{V_{n+1,K_n+1}}{V_{n,K_n}} P^*(A) + \frac{V_{n+1,K_n}}{V_{n,K_n}} \sum_{i=1}^{K_n} (N_i - \sigma)\, \delta_{X_i^*}(A), \quad (\circ)$$

where $\{V_{n,j} : \; n \geq 1, \; 1 \leq j \leq n\}$ is a set of weights which satisfy the recursion

$$V_{n,j} = (n - j\sigma) V_{n+1,j} + V_{n+1,j+1}. \qquad (\Diamond)$$

$\Longrightarrow$ completely characterized by choice of $\sigma < 1$ and a set of weights $V_{n,j}$'s.

Complete predictive structure

$\tilde{P}$ is a Gibbs-type random probability measure of order $\sigma \in (-\infty, 1)$ if and only if it gives rise to predictive distributions of the form

$$\mathbb{P}\left[X_{n+1} \in A \,\middle|\, X^{(n)}\right] = \frac{V_{n+1,K_n+1}}{V_{n,K_n}} P^*(A) + \frac{V_{n+1,K_n}}{V_{n,K_n}} \sum_{i=1}^{K_n} (N_i - \sigma)\,\delta_{X_i^*}(A), \quad (\circ)$$

where $\{V_{n,j} :\ n \geq 1,\ 1 \leq j \leq n\}$ is a set of weights which satisfy the recursion

$$V_{n,j} = (n - j\sigma) V_{n+1,j} + V_{n+1,j+1}. \qquad (\Diamond)$$

$\Longrightarrow$ completely characterized by choice of $\sigma < 1$ and a set of weights $V_{n,j}$'s.

E.g., if $V_{n,j} = \frac{\prod_{i=1}^{k-1}(\theta + i\sigma)}{(\theta+1)_{n-1}}$ with $\sigma \geq 0$ and $\theta > -\sigma$ or $\sigma < 0$ and $\theta = r|\sigma|$ with $r \in \mathbb{N}$, one obtains the two parameter Poisson–Dirichlet (PD) process (Perman, Pitman & Yor, 1992) aka Pitman–Yor process, which yields

$$\mathbb{P}\left[X_{n+1} \in A \,\middle|\, X^{(n)}\right] = \frac{\theta + K_n\sigma}{\theta + n} P^*(A) + \frac{1}{\theta + n} \sum_{i=1}^{K_n} (N_i - \sigma)\delta_{X_i^*}(A).$$

$\Longrightarrow$ if $\sigma = 0$, the PD reduces to the Dirichlet process and $\frac{\theta + K_n\sigma}{\theta + n}$ to $\frac{\theta}{\theta + n}$.

The Gibbs–structure allows to look at the predictive distributions as the result of two steps:

(1) X_{n+1} is a new species with probability

$$V_{n+1,K_n+1}/V_{n,K_n},$$

whereas it equals one of the "old" $\{X_1^*, \ldots, X_{K_n}^*\}$ with probability

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

$\Longrightarrow$ This step depends on n and K_n but not on the frequencies $\mathbf{N}_n = (N_1, \ldots, N_{K_n})$.

The Gibbs–structure allows to look at the predictive distributions as the result of two steps:

(1) X_{n+1} is a new species with probability

$$V_{n+1,K_n+1}/V_{n,K_n},$$

whereas it equals one of the "old" $\{X_1^*,\dots,X_{K_n}^*\}$ with probability

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

$\Longrightarrow$ This step depends on n and K_n but not on the frequencies $\mathbf{N}_n = (N_1,\dots,N_{K_n})$.

(2) (i) Given X_{n+1} is new, it is independently sampled from P^*.

(ii) Given X_{n+1} is a tie, it coincides with X_i^* with probability

$$(N_i - \sigma)/(n - K_n\sigma).$$

Who are the members of this class of priors?

Gnedin and Pitman (2006) provided also a characterization of Gibbs–type priors according to the value of σ:

- $\sigma = 0 \implies$ Dirichlet process or Dirichlet process mixed over its total mass parameter $\theta > 0$;

Who are the members of this class of priors?

Gnedin and Pitman (2006) provided also a characterization of Gibbs–type priors according to the value of σ:

- $\sigma = 0 \implies$ Dirichlet process or Dirichlet process mixed over its total mass parameter $\theta > 0$;
- $0 < \sigma < 1 \implies$ random probability measures closely related to a normalized σ–stable process (Poisson–Kingman models based on the σ-stable process) characterized by σ and a probability distribution γ.

  Special cases: in addition to the PD process another noteworthy example is given by the normalized generalized gamma process (NGG) for which

$$V_{n,j} = \frac{\mathrm{e}^{\beta}\, \sigma^{j-1}}{\Gamma(n)} \sum_{i=0}^{n-1} \binom{n-1}{i} (-1)^i \, \beta^{i/\sigma} \, \Gamma\left(j - \frac{i}{\sigma};\, \beta\right),$$

  where $\beta > 0$, $\sigma \in (0,1)$ and $\Gamma(x, a)$ denotes the incomplete gamma function. If $\sigma = 1/2$ it reduces to the normalized inverse Gaussian process (N–IG).

- $\sigma < 0 \implies$ mixtures of symmetric k–variate Dirichlet distributions

$$(\tilde{p}_1, \ldots, \tilde{p}_K) \sim \text{Dirichlet}(|\sigma|, \ldots, |\sigma|) \qquad (*)$$
$$K \sim \pi(\cdot)$$

- $\sigma < 0 \implies$ mixtures of symmetric k–variate Dirichlet distributions

$$(\tilde{p}_1, \dots, \tilde{p}_K) \sim \text{Dirichlet}(|\sigma|, \dots, |\sigma|) \qquad (*)$$
$$K \sim \pi(\cdot)$$

Special cases:

- If π is degenerate on $r \in \mathbb{N}$ one has symmetric r–variate Dirichlet distributions which corresponds to a PD process with $\sigma < 0$ and $\theta = r|\sigma|$ and is aka Wright–Fisher model.
- The model of Gnedin (2010) arises if, for $r = 1, 2, \dots$ with $\gamma \in (0, 1)$,

$$\pi(r) = \frac{\gamma(1-\gamma)_{r-1}}{r!}$$

- Other interesting cases arise if π is a Poisson distribution (restricted to the positive integers) or a geometric distribution.

- $\sigma < 0 \implies$ mixtures of symmetric k–variate Dirichlet distributions

$$(\tilde{p}_1, \ldots, \tilde{p}_K) \sim \text{Dirichlet}(|\sigma|, \ldots, |\sigma|) \qquad (*)$$
$$K \sim \pi(\cdot)$$

Special cases:

- If π is degenerate on $r \in \mathbb{N}$ one has symmetric r–variate Dirichlet distributions which corresponds to a PD process with $\sigma < 0$ and $\theta = r|\sigma|$ and is aka Wright–Fisher model.
- The model of Gnedin (2010) arises if, for $r = 1, 2, \ldots$ with $\gamma \in (0, 1)$,

$$\pi(r) = \frac{\gamma(1-\gamma)_{r-1}}{r!}$$

- Other interesting cases arise if π is a Poisson distribution (restricted to the positive integers) or a geometric distribution.

Remark.

- If $\sigma \geq 0$ the model assumes the existence of an infinite number of species
- If $\sigma < 0$ (and π not degenerate) the model assumes a random but finite number of species. Interestingly, in Gnedin's model it will have infinite mean!

Full weak support property of Gibbs–type priors

Henceforth focus on:
Gibbs–type priors whose realizations are discrete distributions where the number of support points is not bounded $\iff$ $\sigma \geq 0$ or $\sigma < 0$ with π in $(*)$ having support $\mathbb{N}$ $\implies$ "genuinely nonparametric priors"

Full weak support property of Gibbs–type priors

Henceforth focus on:
Gibbs–type priors whose realizations are discrete distributions where the number of support points is not bounded $\iff$ $\sigma \geq 0$ or $\sigma < 0$ with π in $(*)$ having support $\mathbb{N}$ $\implies$ "genuinely nonparametric priors"

Let Q be a Gibbs–type prior with prior guess $\mathbb{E}[\tilde{P}] := P^*$ and supp$(P^*) = \mathbb{X}$. Then the topological support of Q coincides with the whole space of probability measures $\mathscr{P}$ that is

$$\mathit{supp}(Q) = \mathscr{P}.$$

$\implies$ Gibbs–type priors have full weak support

Stick–breaking representation of Gibbs–type priors with $\sigma > 0$

Recall that a Gibbs–type prior with $0 < \sigma < 1$ is characterized by σ and a distribution γ.

A Gibbs–type prior $\tilde{P} = \sum_{i=1}^{\infty} \tilde{p}_i \delta_{Z_i}$ with $\sigma > 0$ admits stick–breaking representation of the form

$$\tilde{p}_1 = V_1, \qquad \tilde{p}_i = V_i \prod_{j=1}^{i-1} (1 - V_j) \quad i \geq 2$$

with $(V_i)_{i \geq 1}$ being a sequence of r.v.s such that $V_i | V_1, \ldots, V_{i-1}$ admits density function, for any $i \geq 1$,

$$f(v_i | v_1, \ldots, v_{i-1}) = \frac{\sigma}{\Gamma(1-\sigma)} (v_i \prod_{j=1}^{i-1} (1 - v_j))^{-\sigma}$$
$$\times \frac{\int_0^{+\infty} t^{-i\sigma} f_\sigma(t \prod_{j=1}^{i} (1 - v_j))(f_\sigma(t))^{-1} \gamma(\mathrm{d}t)}{\int_0^{+\infty} t^{-(i-1)\sigma} f_\sigma(t \prod_{j=1}^{i-1} (1 - v_j))(f_\sigma(t))^{-1} \gamma(\mathrm{d}t)} \mathbb{1}_{(0,1)}(v_i)$$

with f_σ denoting the density of a positive stable r.v.

$\Longrightarrow$ Stick–breaking representation with dependent weights!

Special cases

- In the PD case the previous representation reduces to the well–known one with $(V_i)_{i\geq 1}$ a sequence of independent r.v.s

$$V_i \sim \text{Beta}(1-\sigma, \theta + i\sigma)$$

Special cases

- In the PD case the previous representation reduces to the well–known one with $(V_i)_{i\geq 1}$ a sequence of independent r.v.s

$$V_i \sim \text{Beta}(1-\sigma, \theta + i\sigma)$$

- In the N–IG case the dependent weights become completely explicit

$$f(v_i|v_1,\ldots,v_{i-1}) = \frac{\left(\frac{a}{\prod_{j=1}^{i-1}(1-V_j)}\right)^{1/4} (v_i)^{-1/2}(1-v_i)^{-5/4+i/4}}{\sqrt{2\pi}\,\mathsf{K}_{-i/2}\left(\sqrt{\frac{a}{\prod_{j=1}^{i-1}(1-V_j)}}\right)} \times \mathsf{K}_{-\frac{1}{2}-i/2}\left(\sqrt{\frac{\frac{a}{\prod_{j=1}^{i-1}(1-V_j)}}{1-v_i}}\right)\mathbb{I}_{(0,1)}(v_i).$$

which can also be represented as $U_i/(U_i+W_i)$ with U_i a generalized inverse Gaussian r.v. (with parameters depending on V_{i-1}) and W_i a positive stable r.v.

Induced distribution on number of clusters

An alternative definition of Gibbs–type priors is as species sampling models (i.e. discrete nonparametric priors $\sum_{i\geq 1} \tilde{p}_i \delta_{Y_i}(\,\cdot\,)$ in which the weights p_i's and locations Y_i are independent) which induce a random partition of the form

$$\Pi_k^n(n_1, \ldots, n_j) = V_{n,j} \prod_{i=1}^{j} (1-\sigma)_{n_i - 1} \qquad (\triangle)$$

for any $n \geq 1$, $j \leq n$ and positive integers $n_1, \ldots, n_j$ such that $\sum_{i=1}^{j} n_i = n$, where $\sigma < 1$ and the $V_{n,j}$'s satisfy the recursion $(\Diamond)$.

Intepretation of $(\triangle)$: probability of observing a specific sample $X_1, \ldots, X_n$ featuring j distinct observations with frequencies $n_1, \ldots, n_j \implies$ exchangeable partition probability function (EPPF), a concept introduced in Pitman (1995).

Induced distribution on number of clusters

An alternative definition of Gibbs–type priors is as species sampling models (i.e. discrete nonparametric priors $\sum_{i\geq 1} \tilde{p}_i \delta_{Y_i}(\,\cdot\,)$ in which the weights p_i's and locations Y_i are independent) which induce a random partition of the form

$$\Pi_k^n(n_1, \ldots, n_j) = V_{n,j} \prod_{i=1}^{j} (1-\sigma)_{n_i - 1} \qquad (\triangle)$$

for any $n \geq 1$, $j \leq n$ and positive integers $n_1, \ldots, n_j$ such that $\sum_{i=1}^{j} n_i = n$, where $\sigma < 1$ and the $V_{n,j}$'s satisfy the recursion $(\Diamond)$.

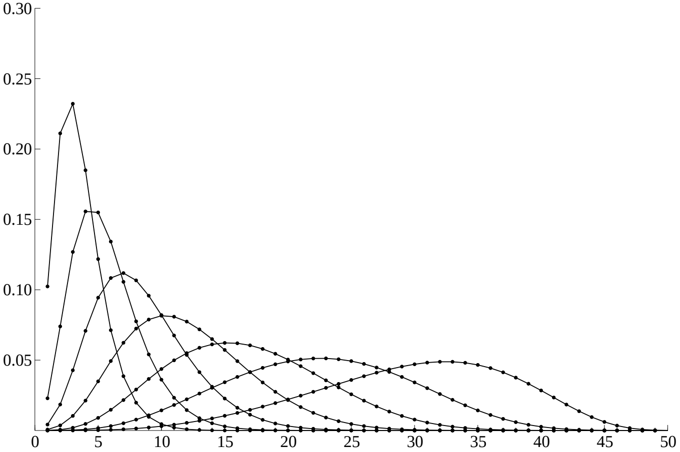
Intepretation of $(\triangle)$: probability of observing a specific sample $X_1, \ldots, X_n$ featuring j distinct observations with frequencies $n_1, \ldots, n_j$ $\implies$ exchangeable partition probability function (EPPF), a concept introduced in Pitman (1995).

Consequently, one obtains the (prior) distribution of the number of clusters by summing over all possible partitions of a given size

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

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

Prior distribution of the number of clusters as σ varies



Prior distributions on the number of groups corresponding to a NGG process with $n = 50, \beta = 1$ and $\sigma = 0.1, 0.2, 0.3, \ldots, 0.8$ (from left to right).

In general, the dependence of the distribution of K_n on the prior parameters is as follows:

- σ controls the "flatness" (or variability) of the (prior) distribution of K_n.
- the possible second parameter (θ in the PD and β in the NGG case) controls the location of the (prior) distribution of K_n

In general, the dependence of the distribution of K_n on the prior parameters is as follows:

- σ controls the "flatness" (or variability) of the (prior) distribution of K_n.
- the possible second parameter (θ in the PD and β in the NGG case) controls the location of the (prior) distribution of K_n

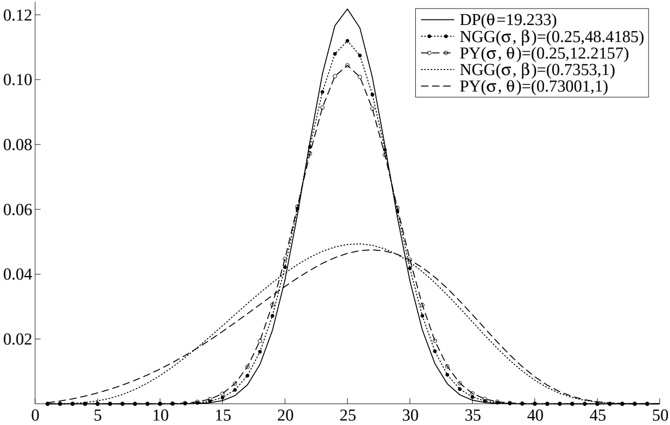
Comparative example of different Gibbs–type priors:

- $n = 50$ and the prior expected number of clusters is 25 $\Longrightarrow$ fix the prior parameters s.t. $\mathbb{E}(K_{50}) = 25$.
- 5 different models:
  - Dirichlet process with $\theta = 19.233$;
  - PD processes with $(\sigma, \theta) = (0.73001, 1)$ and $(\sigma, \theta) = (0.25, 12.2157)$;
  - NGG processes with $(\sigma, \beta) = (0.7353, 1)$ and $(0.25, 48.4185)$.

$\Longrightarrow$ Dirichlet process implies a highly peaked distribution of K_n:

- circumvented by placing a prior on θ; though would such a prior (and its parameters) be the same for whatever sample size?
- moreover, why one should add another layer to the model which can be avoided by selecting a slightly more general process?

Prior distribution of the number of clusters



Prior distributions on the number of clusters corresponding to the Dirichlet, the PD and the NGG processes. The values of the parameters are set in such a way that $\mathbb{E}(K_{50}) = 25$.

Toy mixture example

- $n = 50$ observations are drawn from a uniform mixture of two well-separated Gaussian distributions, $\mathrm{N}(1, 0.2)$ and $\mathrm{N}(10, 0.2)$;
- nonparametric mixture model

$$
\begin{aligned}
(Y_i \mid \mathrm{m}_i, \mathrm{v}_i) &\overset{\text{ind}}{\sim} \mathrm{N}(\mathrm{m}_i, \mathrm{v}_i), & i = 1, \ldots, n \\
(\mathrm{m}_i, \mathrm{v}_i \mid \tilde{p}) &\overset{\text{iid}}{\sim} \tilde{p} & i = 1, \ldots, n \\
\tilde{p} &\sim Q
\end{aligned}
$$

with Q a Gibbs–type prior and standard specifications for P^*;

Toy mixture example

- $n = 50$ observations are drawn from a uniform mixture of two well-separated Gaussian distributions, $\mathrm{N}(1, 0.2)$ and $\mathrm{N}(10, 0.2)$;
- nonparametric mixture model

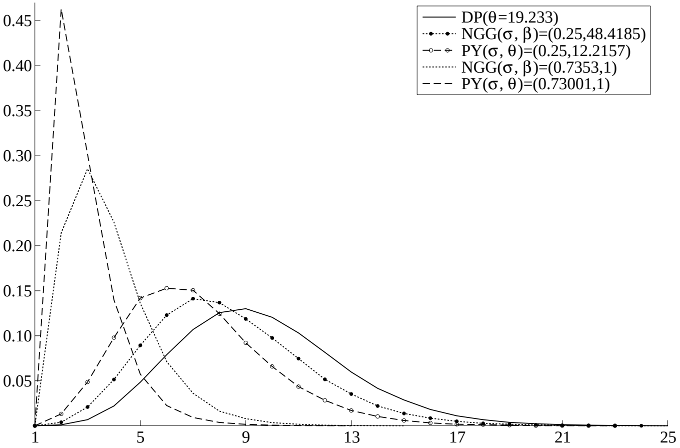
$$(Y_i \mid \mathrm{m}_i, \mathrm{v}_i) \overset{\text{ind}}{\sim} \mathrm{N}(\mathrm{m}_i, \mathrm{v}_i), \qquad i = 1, \ldots, n$$
$$(\mathrm{m}_i, \mathrm{v}_i \mid \tilde{p}) \overset{\text{iid}}{\sim} \tilde{p} \qquad i = 1, \ldots, n$$
$$\tilde{p} \sim Q$$

  with Q a Gibbs–type prior and standard specifications for P^*;
- As Q we consider the previous 5 priors (chosen so that $E(K_{50}) = 25$), which in this case correspond to a prior opinion on K_{50} remarkably far from the true number of components, namely 2.

Are the models flexible enough to shift a posteriori towards the correct number of components?

$\Longrightarrow$ the larger σ the better is the posterior estimate of K_n.

Posterior distribution of the number of clusters



Posterior distributions on the number of groups corresponding to various choices of Gibbs–type priors with $n = 50$ *and* $\mathbb{E}(K_{50}) = 25$*.*

Data structure in species sampling problems

- $X^{(n)}$ = basic sample of draws from a population containing different species (plants, genes, animals,...). Information:
  - ◇ sample size n and number of distinct species in the sample K_n;
  - ◇ a collection of frequencies $\mathbf{N} = (N_1, \ldots, N_{K_n})$ s.t. $\sum_{i=1}^{K_n} N_i = n$;
  - ◇ the labels (names) X_i^*'s of the distinct species, for $i = 1, \ldots, K_n$.

Data structure in species sampling problems

- $X^{(n)}$ = basic sample of draws from a population containing different species (plants, genes, animals,...). Information:
  - ◇ sample size n and number of distinct species in the sample K_n;
  - ◇ a collection of frequencies $\mathbf{N} = (N_1, \dots, N_{K_n})$ s.t. $\sum_{i=1}^{K_n} N_i = n$;
  - ◇ the labels (names) X_i^*'s of the distinct species, for $i = 1, \dots, K_n$.
- The information provided by $\mathbf{N}$ can also be coded by $\mathbf{M} := (M_1, \dots, M_n)$

  $$M_i = \text{number of species in the sample } X^{(n)} \text{ having frequency } i.$$

  Note that $\sum_{i=1}^{n} M_{i,n} = K_n$ and $\sum_{i=1}^{n} iM_{i,n} = n$.
- Example: Consider a basic sample such that
  - ◇ $n = 10$ with $j = 4$ and frequencies $(n_1, n_2, n_3, n_4) = (2, 5, 2, 1)$.
  - ◇ equivalently we can code this information as

  $$(m_1, m_2, \dots, m_{10}) = (1, 2, 0, 0, 1, \dots, 0),$$

  meaning that 1 species appears once, 2 appear twice and 1 five times.

Prediction problems

Given the basic sample $X^{(n)}$, the inferential goal consists in prediction about various features of an additional sample $X^{(m)} := (X_{n+1}, \ldots, X_{n+m})$.

Discovery probability $\Longrightarrow$ estimation of

1. the probability of discovering at the (n+1)–th sampling step either a new species or an "old" species with frequency r;
2. the probability of discovering at the (n+m+1)–th step either a new species or an "old" species with frequency r without observing $X^{(m)}$.

Prediction problems

Given the basic sample $X^{(n)}$, the inferential goal consists in prediction about various features of an additional sample $X^{(m)} := (X_{n+1}, \ldots, X_{n+m})$.

Discovery probability $\Longrightarrow$ estimation of

1. the probability of discovering at the (n+1)–th sampling step either a new species or an "old" species with frequency r;
2. the probability of discovering at the (n+m+1)–th step either a new species or an "old" species with frequency r without observing $X^{(m)}$.

Remark. These can be, in turn, used to obtain straightforward estimates of:

- the discovery probability for rare species i.e. the probability of discovering a species which is either new or has frequency at most τ at the (n+m+1)–th step $\Longrightarrow$ rare species estimation
- an optimal additional sample size: sampling is stopped once the probability of sampling new or rare species is below a certain threshold
- the sample coverage, i.e. the proportion of species in the population detected in the basic sample $X^{(n)}$ or in an enlarged sample $X^{(n+m)}$.

Frequentist nonparametric estimators

- Turing estimator (Good, 1953; Mao & Lindsay, 2002): probability of discovering a species with frequency r in $X^{(n)}$ at (n+1)–th step is

$$(r+1)\frac{m_{r+1}}{n} \qquad (\star)$$

  and for $r = 0$ one obtains the discovery probability of a new species $\frac{m_1}{n}$.

  $\Longrightarrow$ depends on m_{r+1} (number of species with frequency $r+1$): counterintuitive! It should be based on m_r. E.g. if $m_{r+1} = 0$, the estimated probability of detecting a species with frequency r would be 0.

- Good–Toulmin estimator (Good & Toulmin, 1956; Mao, 2004): estimator for the probability of discovering a new species at (n+m+1)–th step.

  $\Longrightarrow$ unstable if the size of the additional unobserved sample m is larger than n (estimated probability becomes either < 0 or > 1).

- No frequentist nonparametric estimator for the probability of discovering a species with frequency r at (n+m+1)–th sampling step is available.

BNP approach to discovery probability estimation

We assume the data $(X_n)_{n\geq 1}$ are exchangeable and a Gibbs–type prior as corresponding de Finetti measure. The resulting estimators are as follows:

- BNP analog to Turing estimator: probability of discovering a species with frequency r in $X^{(n)}$ at the (n+1)–th sampling step

$$\mathbb{P}[X_{n+1} = \text{species with frequency } r \mid X^{(n)}] = \frac{V_{n+1,k}(r-\sigma)}{V_{n,k}}\, m_r,$$

and the discovery probability of a new species

$$\mathbb{P}[X_{n+1} = \text{``new''} \mid X^{(n)}] = \frac{V_{n+1,k+1}}{V_{n,k}}.$$

Remark 1. Probability of sampling a species with frequency r depends, in agreement with intuition, on m_r and also on $K_n = k$.

- BNP analog of the Good–Toulmin estimator: estimator for the probability of discovering a new species at the (n+m+1)–th step

$$\mathbb{P}[X_{n+m+1} = \text{“new”} \mid X^{(n)}] = \sum_{j=0}^{m} \frac{V_{n+m+1,k+j+1}}{V_{n,k}} \frac{\mathscr{C}(m,j;\sigma,-n+k\sigma)}{\sigma^j}$$

  with $\mathscr{C}(m,j;\sigma,-n+k\sigma) = j!^{-1} \sum_{l=0}^{j}(-1)^l \binom{j}{l} (n-\sigma(l+k))_m$ being the non–central generalized factorial coefficient.

- BNP estimator for the probability of discovering a species with frequency r at the (n+m+1)–th sampling step

$$\mathbb{P}[X_{n+m+1} = \text{species with frequency } r \mid X^{(n)}]$$

  is available in closed form and yields immediately an estimator of the rare species discovery probability.

The discovery probability in the PD process case

The natural candidate for applications is the PD process which yields completely explicit estimators.

Remark. The Dirichlet process is not appropriate for conceptual reasons and also because it lacks the required flexibility in modeling the growth rate by imposing a logarithmic growth of new species, where the PD process allows for rates n^{σ} for $\sigma \in (0, 1)$. See also Teh (2006).

The discovery probability in the PD process case

The natural candidate for applications is the PD process which yields completely explicit estimators.

Remark. The Dirichlet process is not appropriate for conceptual reasons and also because it lacks the required flexibility in modeling the growth rate by imposing a logarithmic growth of new species, where the PD process allows for rates n^σ for $\sigma \in (0,1)$. See also Teh (2006).

- PD analog to Turing estimator: probability of discovering a species with frequency r in $X^{(n)}$ at the (n+1)–th sampling step is given by

$$\mathbb{P}[X_{n+1} = \text{species with frequency } r \mid X^{(n)}] = \frac{r-\sigma}{\theta+n}\, m_r,$$

and the discovery probability of a new species coincides with

$$\mathbb{P}[X_{n+1} = \text{“new”} \mid X^{(n)}] = \frac{\theta+\sigma k}{\theta+n}.$$

- PD analog of the Good–Toulmin estimator: estimator for the probability of discovering a new species at the (n+m+1)–th sampling step is

$$\mathbb{P}[X_{n+m+1} = \text{“new”} \mid X^{(n)}] = \frac{\theta + k\sigma}{\theta + n} \frac{(\theta + n + \sigma)_m}{(\theta + n + 1)_m}$$

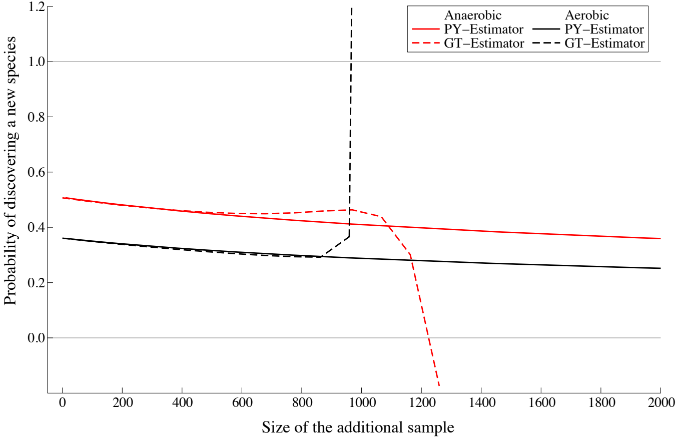
- PD estimator for the probability of discovering a species with frequency r at the (n+m+1)–th step

$$\mathbb{P}[X_{n+m+1} = \text{species with frequency } r \mid X^{(n)}] =$$

$$\sum_{i=1}^{r} m_i (i - \sigma)_{r+1-i} \binom{m}{r-i} \frac{(\theta + n - i + \sigma)_{m-r+i}}{(\theta + n)_{m+1}}$$

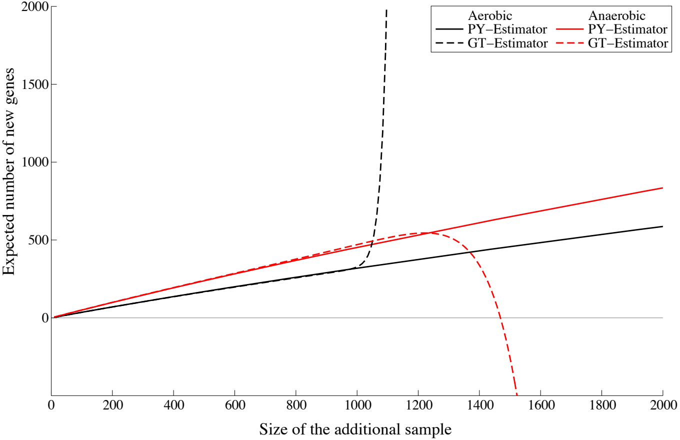
$$+ \frac{(1-\sigma)_r}{(\theta + n)_{m+1}} \left[(\theta + k\sigma)(\theta + n + \sigma)_{m-r} - \prod_{i=k}^{k+m-r} (\theta + i\sigma) \right]$$

Discovery probability in an additional sample of size *m*.



EST data from Naegleria gruberi aerobic and anaerobic cDNA libraries with basic sample $n \cong 950$: Good–Toulmin (GT) and PD process (PD) estimators of the probability of discovering a new gene at the $(n+m+1)$–th sampling step for $m = 1, \ldots, 2000$.

Expected number of new genes in an additional sample of size m.



EST data from Naegleria gruberi aerobic and anaerobic cDNA libraries with basic sample $n \cong 950$: Good–Toulmin (GT) and Pitman–Yor (PY) estimators of the number of new genes to be observed in an additional sample of size $m = 1, \ldots, 2000$.

Some remarks on BNP models for species sampling problems

- BNP estimators available for other quantities of interest in species sampling problems (completely explicit in the PD case).
- BNP models correspond to large probabilistic models in which all objects of potential interest are modeled jointly and coherently thus leading to intuitive predictive structures
  $\Longrightarrow$ avoids ad–hoc procedures and incoherencies sometimes connected with frequentist nonparametric procedures.

Some remarks on BNP models for species sampling problems

- BNP estimators available for other quantities of interest in species sampling problems (completely explicit in the PD case).
- BNP models correspond to large probabilistic models in which all objects of potential interest are modeled jointly and coherently thus leading to intuitive predictive structures
  $\Longrightarrow$ avoids ad–hoc procedures and incoherencies sometimes connected with frequentist nonparametric procedures.
- Gibbs–type priors with $\sigma > 0$ (recall that they assume an infinite number of species) are ideally suited for populations with large unknown number of species $\Longrightarrow$ typical case in Genomics.
- In Ecology "∞" assumption often too strong $\Longrightarrow$ Gibbs–type priors with $\sigma < 0$ (*work in progress* which yields a surprising by–product: by combining Gibbs-type priors with $\sigma > 0$ and $\sigma < 0$ is possible to identify situations in which frequentist estimators work).

Frequentist Posterior Consistency

"What if" or frequentist approach to consistency (Diaconis and Freedman, 1986): What happens if the data are not exchangeable but i.i.d. from a "true" P_0? Does the posterior $Q(\cdot|X^{(n)})$ accumulate around P_0 as the sample size increases?

Q is weakly consistent at P_0 if for every A_ε

$$Q(A_\varepsilon|X^{(n)}) \overset{n\to\infty}{\longrightarrow} 1 \qquad a.s. - P_0^\infty$$

with A_ε a weak neighbourhood of P_0 and P_0^∞ the infinite product measure.

Frequentist Posterior Consistency

"What if" or frequentist approach to consistency (Diaconis and Freedman, 1986): What happens if the data are not exchangeable but i.i.d. from a "true" P_0? Does the posterior $Q(\,\cdot\,|X^{(n)})$ accumulate around P_0 as the sample size increases?

Q is weakly consistent at P_0 if for every A_ε

$$Q(A_\varepsilon|X^{(n)}) \stackrel{n\to\infty}{\longrightarrow} 1 \qquad a.s. - P_0^\infty$$

with A_ε a weak neighbourhood of P_0 and P_0^∞ the infinite product measure.

We investigate consistency for Gibbs–type priors with $\sigma \in (-\infty, 0)$
Proof strategy consists in showing that

- $\mathbb{E}[\tilde{P} \mid X^{(n)}] \stackrel{n\to\infty}{\longrightarrow} P_0$ a.s.–P_0^∞ $\Longleftrightarrow$ by the predictive structure ($\circ$) of Gibbs–type priors: $\mathbb{P}[X_{n+1} = \text{"new"} \mid X^{(n)}] = V_{n+1,k+1}/V_{n,k} \stackrel{n\to\infty}{\longrightarrow} 0$ a.s.–P_0^∞
- $\text{Var}[\tilde{P} \mid X^{(n)}] \stackrel{n\to\infty}{\longrightarrow} 0$ a.s.–P_0^∞ by finding a suitable bound on the variance.

The case of discrete "true" data generating distribution P_0

Two cases according to the type of "true" data generating distribution P_0:

- P_0 is discrete (with either finite or infinite support points)
- P_0 is diffuse (i.e. $P_0(\{x\}) = 0$ for every $x \in \mathbb{X}$ termed "continuous")

The case of discrete "true" data generating distribution P_0

Two cases according to the type of "true" data generating distribution P_0:

- P_0 is discrete (with either finite or infinite support points)
- P_0 is diffuse (i.e. $P_0(\{x\}) = 0$ for every $x \in \mathbb{X}$ termed "continuous")

Let Q be a Gibbs–type prior with $\sigma < 0$ and P_0 a discrete "true" distribution. Then, under an extremely mild technical condition, Q is consistent at P_0.

Remark. The technical condition serves only for pinning down the proof in general: one can comfortably speak of having "essentially always" consistency (for not covered instances consistency shown case-by-case).

The case of discrete "true" data generating distribution P_0

Two cases according to the type of "true" data generating distribution P_0:

- P_0 is discrete (with either finite or infinite support points)
- P_0 is diffuse (i.e. $P_0(\{x\}) = 0$ for every $x \in \mathbb{X}$ termed "continuous")

Let Q be a Gibbs–type prior with $\sigma < 0$ and P_0 a discrete "true" distribution. Then, under an extremely mild technical condition, Q is consistent at P_0.

Remark. The technical condition serves only for pinning down the proof in general: one can comfortably speak of having "essentially always" consistency (for not covered instances consistency shown case-by-case).

$\Longrightarrow$ frequentist consistency is guaranteed when modeling data coming from a discrete distribution like in species sampling problems

$\Updownarrow$

Discrete nonparametric priors are consistent
for data generated by discrete distributions.

The case of continuous "true" data generating distribution P_0

Discrete $P_0 \implies$ consistency "essentially always"

Contin. $P_0 \implies$ wide range of asymptotic beahviours including erratic ones.

Remark. Since P_0 is continuous, the number of distinct observations in a sample of size n, K_n, is precisely n. Also recall that Gibbs–type priors with $\sigma < 0$ are mixtures of symmetric Dirichlet distributions

$$(\tilde{p}_1, \ldots, \tilde{p}_K) \sim \text{Dirichlet}(|\sigma|, \ldots, |\sigma|)$$
$$K \sim \pi(\cdot)$$

The case of continuous "true" data generating distribution P_0

Discrete $P_0 \implies$ consistency "essentially always"

Contin. $P_0 \implies$ wide range of asymptotic beahviours including erratic ones.

Remark. Since P_0 is continuous, the number of distinct observations in a sample of size n, K_n, is precisely n. Also recall that Gibbs–type priors with $\sigma < 0$ are mixtures of symmetric Dirichlet distributions

$$(\tilde{p}_1, \ldots, \tilde{p}_K) \sim \text{Dirichlet}(|\sigma|, \ldots, |\sigma|)$$
$$K \sim \pi(\cdot)$$

Example 1: Gibbs–type prior with $\sigma = -1$ with Poisson(λ) mixing distribution π (restricted to the positive integers).
Key quantity is the probability of obtaining a new observation:

$$\mathbb{P}[X_{n+1} = \text{"new"} \mid X^{(n)}] = V_{n+1,n+1}/V_{n,n}$$
$$= \frac{\lambda n}{(2n+1)(2n)} \frac{{}_1F_1(n; 2n; \lambda)}{{}_1F_1(n+1; 2n+2; \lambda)} \sim \frac{\lambda}{2(2n+1)} \xrightarrow{n\to\infty} 0$$

This, combined with some other arguments, shows that such a prior is consistent at any continuous P_0.

Example 2: Gnedin's model with $\sigma = -1$ and parameter $\gamma \in (0, 1)$.
For continuous P_0 we obtain:

$$\mathbb{P}[X_{n+1} = \text{“new”} \mid X^{(n)}] = V_{n+1,n+1}/V_{n,n} = \frac{n(n-\gamma)}{n(\gamma+n)} \xrightarrow{n\to\infty} 1$$

This, combined with some other arguments, shows that Q is inconsistent at any continuous P_0. Moreover, not only it is inconsistent: it concentrates around the prior guess P^* meaning that no learning at all takes place $\Longrightarrow$ "total" inconsistency.

Example 2: Gnedin's model with $\sigma = -1$ and parameter $\gamma \in (0, 1)$.
For continuous P_0 we obtain:

$$\mathbb{P}[X_{n+1} = \text{"new"} \mid X^{(n)}] = V_{n+1,n+1}/V_{n,n} = \frac{n(n-\gamma)}{n(\gamma+n)} \xrightarrow{n\to\infty} 1$$

This, combined with some other arguments, shows that Q is inconsistent at any continuous P_0. Moreover, not only it is inconsistent: it concentrates around the prior guess P^* meaning that no learning at all takes place $\Longrightarrow$ "total" inconsistency.

Example 3: Gibbs–type prior with $\sigma = -1$ and geometric(η) mixing dist. π.
For continuous P_0 we obtain:

$$\mathbb{P}[X_{n+1} = \text{"new"} \mid X^{(n)}] = V_{n+1,n+1}/V_{n,n}$$

$$= \frac{\eta n(n+1)}{(2n+1)(2n)} \frac{{}_2F_1(n, n+1; 2n; \eta)}{{}_2F_1(n+1, n+2; 2n+2; \eta)} \xrightarrow{n\to\infty} \frac{2-\eta-2\sqrt{1-\eta}}{\eta} \in [0, 1]$$

$\Longrightarrow$ the posterior concentrates on $\alpha P^* + (1-\alpha)P_0$ with $\alpha = \frac{2-\eta-2\sqrt{1-\eta}}{\eta}$:
therefore, by tuning the parameter η, one can obtain any possible posterior behaviour ranging from consistency ($\eta = 0$) to "total" inconsistency ($\eta = 1$).

The general consistency result for continuous P_0 is then as follows:

Let Q be a Gibbs–type prior with $\sigma < 0$ and P_0 a continuous "true" distribution. Then, Q is consistent at P_0 provided for sufficiently large x and for some $M < \infty$

$$\frac{\pi(x+1)}{\pi(x)} \leq \frac{M}{x}. \qquad (\triangledown)$$

$\Longrightarrow$ $(\triangledown)$ requires the tail of π to be sufficiently light and is close to necessary.

The general consistency result for continuous P_0 is then as follows:

> *Let Q be a Gibbs–type prior with $\sigma < 0$ and P_0 a continuous "true" distribution. Then, Q is consistent at P_0 provided for sufficiently large x and for some $M < \infty$*
>
> $$\frac{\pi(x+1)}{\pi(x)} \leq \frac{M}{x}. \qquad (\triangledown)$$

$\Longrightarrow$ $(\triangledown)$ requires the tail of π to be sufficiently light and is close to necessary.

Remark. The "extremely mild" technical condition for the case of discrete P_0 corresponds to asking π to be ultimately decreasing.

What does this asymptotic analysis tell us?

Practical level: Neat conditions which guarantee consistency for a large class of nonparametric priors increasingly used in practice.

Foundational level: discrete $\tilde{P}$ designed to model discrete distrib. and should not be used to model data from continuous distributions.

What does this asymptotic analysis tell us?

Practical level: Neat conditions which guarantee consistency for a large class of nonparametric priors increasingly used in practice.

Foundational level: discrete $\tilde{P}$ designed to model discrete distrib. and should not be used to model data from continuous distributions.

Remark. Dirichlet process enjoys:
◇ full weak support property
◇ weak consistency for continuous $P_0 \implies$ misleading!

But as the sample size n diverges:
◇ P_0 generates $(X_n)_{n\geq 1}$ containing no ties with probability 1
◇ a discrete $\tilde{P}$ generates $(X_n)_{n\geq 1}$ containing no ties with probability 0
$\implies$ model and data generating mechanism are incompatible!

What does this asymptotic analysis tell us?

Practical level: Neat conditions which guarantee consistency for a large class of nonparametric priors increasingly used in practice.

Foundational level: discrete $\tilde{P}$ designed to model discrete distrib. and should not be used to model data from continuous distributions.

Remark. Dirichlet process enjoys:
◇ full weak support property
◇ weak consistency for continuous $P_0 \implies$ misleading!

But as the sample size n diverges:
◇ P_0 generates $(X_n)_{n \geq 1}$ containing no ties with probability 1
◇ a discrete $\tilde{P}$ generates $(X_n)_{n \geq 1}$ containing no ties with probability 0
$\implies$ model and data generating mechanism are incompatible!

For discrete Q it is:
◇ irrelevant to be consistent at continuous P_0 (it is just a coincidence if they are e.g. Dirichlet, Gibbs with Poisson mixing);
◇ important to be consistent at discrete P_0 and they are!

References

- De Blasi, Lijoi, & Prünster (2012). An asymptotic analysis of a class of discrete nonparametric priors. Tech. Report.
- Diaconis & Freedman (1986). On the consistency of Bayes estimates. *Ann. Statist.* **14**, 1–26.
- Gnedin (2010). A species sampling model with finitely many types. *Elect. Comm. Probab.* **15**, 79–88.
- Gnedin & Pitman (2006). Exchangeable Gibbs partitions and Stirling triangles. *J. Math. Sci. (N.Y.)* **138**, 5674–5685.
- Good & Toulmin (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* **43**, 45–63.
- Good (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40**, 237–64.
- Favaro, Lijoi & Prünster (2012). On the stick–breaking representation of normalized inverse Gaussian priors. Biometrika **99**, 663-674.
- Favaro, Lijoi & Prünster (2012). A new estimator of the discovery probability. *Biometrics*, in press.
- Ferguson (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209–30.
- Ferguson (1974). Prior distributions on spaces of probability measures. *Ann. Statist.* **2**, 615–29.
- Lo (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Ann. Statist.* **12**, 351–357.
- Mao & Lindsay (2002). A Poisson model for the coverage problem with a genomic application. *Biometrika* **89**, 669–681.
- Mao (2004). Prediction of the conditional probability of discovering a new class. *J. Am. Statist. Assoc.* **99**, 1108–1118.
- Perman, Pitman & Yor (1992). Size-biased sampling of Poisson point processes and excursions. *Probab. Theory Related Fields* **92**, 21–39.
- Teh (2006). A Hierarchical Bayesian Language Model based on Pitman-Yor Processes. *Coling/ACL 2006*, 985-992.