Discrete Random Structures in Bayesian Nonparametrics

I: The Predictive approach and species sampling

Igor Prünster

Bocconi University

Outline

Induction & Exchangeability

Original Problem of Induction as described in Hume (1739)

> *Is there a rational justification of inductive methods i.e. of methods that predict or infer that "instances of which we have had no experience resemble those of which we have had experience"? We tend to believe that things behave in a regular manner i.e. that patterns in the behaviour of objects will persist into the future (sometimes called Principle of the Uniformity of Nature)*

Hume's negative answer: Induction cannot be proved deductively (it is contingent!) nor inductively (it would be a tautology!)

$\Longrightarrow$ Nonetheless induction or "going beyond the present testimony of the senses and the records of our memory" is essential in scientific reasoning as well as in everyday life.

A probabilistic approach to induction: Exchangeability

> *"In the philosophical arena, the problem of induction, its meaning, use and justification, has given rise to endless controversy, which, in the absence of an appropriate probabilistic framework, has inevitably been fruitless, leaving the major issues unresolved."* [de Finetti, 1975]

$\Longrightarrow$ de Finetti's pragmatic point of view: reformulate the uniformity principle in probabilistic terms by means of the concept of exchangeability, which requires probability statements concerning a set of observations to be invariant with respect to their order.

Exchangeability

$(X_n)_{n\geq 1}$ is exchangeable if for every finite permutation π

$$(X_1, X_2, \ldots, X_n) \stackrel{d}{=} (X_{\pi(1)}, X_{\pi(2)}, \ldots, X_{\pi(n)}).$$

According to de Finetti prediction is the ultimate goal

> *"science cannot limit itself to theorize about accomplished facts but must foresee."* [de Finetti, 1931]

Given the exchangeability assumption, prediction reduces to computing conditional probabilities.

de Finetti's representation theorem

A sequence of $\mathbb{X}$–valued observations $(X_n)_{n\geq 1}$ is exchangeable if and only if

$$\mathbb{P}[X_1 \in A_1, \ldots, X_n \in A_n] = \int_{\mathscr{P}} \prod_{i=1}^{n} P(A_i) Q(\mathrm{d}P)$$

for any $n \geq 1$, where $\mathscr{P}$ is the space of probability measures on $\mathbb{X}$.

$\Longrightarrow$ Q 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 $\tilde{P}$.

Remark: The representation theorem provides a neat justification of the use of prior distributions. It is the assumption of exchangeability (the Bayesian analog to the frequentist i.i.d. assumption) that implies the existence of a prior distribution. In Diaconis' words "a philosophically sensational result". See also Regazzini & Petris (1992) for interesting considerations.

Equivalently one can state the representation theorem in hierarchical form as

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

Prediction

Carnap (1950) provides a taxonomy of the varieties of inductive inference and, in particular:

> *Predictive inference, which is defined as inference from one sample to another sample not overlapping the first, is "the most important and fundamental kind of inductive inference". It includes the special case, known as singular predictive inference, in which the second sample consists of just one individual.*

Within de Finetti's framework one has:

- Singular (or one–step) prediction

$$\underbrace{\mathbb{P}[X_{n+1} \in A \,|X^{(n)}]}_{\text{predicitve distribution} \in \mathscr{P}} = \int_{\mathscr{P}} P(A) \underbrace{Q(\mathrm{d}P|X^{(n)})}_{\text{posterior distribution}}$$

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

- m–step prediction

$$\mathbb{P}[X_{n+1} \in A_1, \ldots, X_{n+m} \in A_m|X^{(n)}] = \int_{\mathscr{P}} \prod_{i=1}^{m} P(A_i)\, Q(\mathrm{d}P|X^{(n)}).$$

Depending on the structure of the prior Q we have:

- If Q is degenerate on a subclass of $\mathscr{P}$ indexed by a finite dimensional parameter

  $\Longrightarrow$ parametric model

  e.g. $Q\{\text{Gaussian distributions with parameter } (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+\} = 1$

- Otherwise nonparametric model

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

The Bayesian framework was laid out in its full generality during the 30's by de Finetti, but as far as the nonparametric side is concerned, still in 1972 Lindley wrote

> *"It is perhaps worth stopping to remark that the problem is a technical one; the Bayesian method embraces non-parametric problems but cannot solve them because the requisite tool is missing."*

$\Longrightarrow$ Breakthrough in Ferguson (1973) with the introduction of the Dirichlet process prior

Discrete nonparametric priors

If the de Finetti measure 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 any (exchangeable) vector $(X_1, \ldots, X_n)$ generated by Q 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 $\sum_{i=1}^{K_n} N_i = n$.

Species sampling: $(\Diamond)$ model for species distribution within a population

- X_i^* is the label of the i–th distinct species in the sample;
- N_i is the frequency of the i–th distinct species;
- K_n is total number of distinct species in the sample.

$\Longrightarrow$ Species metaphor

Dirichlet process from a predictive point of view

Recall the predictive construction of the DP: Assume

- $(X_n)_{n\geq 1}$ is an exchangeable sequence.
- Prior guess at law of any of the X_i's is P^*;
- The strength of the prior belief is measured by a parameter $\theta > 0$.
- Problem: How to predict the distribution of X_{n+1} conditional on a sample $X^{(n)}$ with K_n distinct values $X_1^*, \ldots, X_{K_n}^*$ and frequencies $N_1, \ldots, N_{K_n}$;

Idea: Predict the distribution of X_{n+1} as a linear combination of P^* and the empirical measure $n^{-1}\sum_{i=1}^{K_n} N_i \delta_{X_i^*}$, namely

$$\mathbb{P}[X_{n+1} \in \cdot \mid X^{(n)}] = \underbrace{\frac{\theta}{\theta+n}}_{\mathbb{P}[X_{n+1}=\text{“new”} \mid X^{(n)}]} \underbrace{P^*(\cdot)}_{\text{prior guess}} + \underbrace{\frac{n}{\theta+n}}_{\mathbb{P}[X_{n+1}=\text{“old”} \mid X^{(n)}]} \underbrace{\frac{1}{n}\sum_{i=1}^{K_n} N_i\, \delta_{X_i^*}(\cdot)}_{\text{emprirical measure}}$$

$\Longrightarrow$ Predictive distributions of the Dirichlet process (DP) [Ferguson, 1973].

Remark: The de Finetti measure Q of $(X_n)_{n\geq 1}$ is a DP prior iff the prediction rule is a linear combination of P^* and the empirical measure [Regazzini, 1978; Lo, 1991]

The number of clusters generated by the Dirichlet process

The number of distinct values (species, agents, clusters, components etc.) K_n generated by a sample $X^{(n)}$ is a key quantity in several modeling and inferential contexts. For the DP [Antoniak, 1974] one has

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

with $(\theta)_n := \theta(\theta+1)\dots(\theta+n-1)$ and $|s(n,k)|$ the signless Stirling number of the first kind (and $|s(n,k,r)|$ the non–central version).

m–step prediction: Given a basic sample $X^{(n)}$, prediction of various features of an additional sample $X^{(m)} := (X_{n+1}, \dots, X_{n+m})$ is of interest. In particular,

$$K_m^{(n)} := K_{m+n} - K_n$$

is the number of new species to be recorded in $X^{(m)}$ given $X^{(n)}$.
In the DP case [Favaro, P & Walker, 2011] one has

$$\mathbb{P}\left[K_m^{(n)} = j | X^{(n)}\right] = \frac{\theta^j (\theta)_n}{(\theta)_{(n+m)}} |s(m,j,n)| \quad \mathbb{E}\left[K_m^{(n)} | X^{(n)}\right] = \sum_{i=1}^{m} \frac{\theta}{\theta + n + i - 1}$$

$\Longrightarrow$ The only sample information affecting the distribution of $K_m^{(n)}|X^{(n)}$ is the size n; it depends neither on K_n nor on $N_1, \dots, N_{K_n}$!

Since $\mathbb{P}[X_{n+1} = \text{“new”} \mid X^{(n)}] = \frac{\theta}{\theta+n}$ for any sample size n, clearly K_n will diverge as the sample size $n \to \infty$. But which is its growth rate?

- In the unconditional case, by Korwar and Hollander (1973), one has

$$\frac{K_n}{\log n} \xrightarrow{\text{a.s.}} \theta \qquad (n \to \infty)$$

- In the conditional case, given a fixed sample $X^{(n)}$, one has

$$\frac{K_m^{(n)}}{\log m} \,\Big|\, X^{(n)} \xrightarrow{\text{a.s.}} \theta \qquad (m \to \infty).$$

Remark: Ideally one would like a model with a flexible growth rate which depends on the model parameters (e.g. n^σ with parameter $\sigma \in (0,1)$). Instead, the DP displays a logarithmic growth and such a behaviour is unaffected by the sample $X^{(n)}$ $\implies$ highly restrictive and inappropriate in applications [e.g. linguistics (Teh, 2006), certain bipartite graphs (Caron, 2012), network models (Caron & Fox, 2017) and species sampling]

> *Can such a model be really considered nonparametric? What is responsible for such a restrictive behaviour?*
>
> *It all boils down to* $\mathbb{P}[X_{n+1} = \text{“new”} \mid X^{(n)}] = \frac{\theta}{\theta+n}$ *!*

Probability of discovering a new species

As seen, a key quantity in the analysis of discrete nonparametric priors is the probability of discovering a new species

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

Fundamental Characterization:
Based on $(*)$, discrete $\tilde{P}$ classified in 3 categories [De Blasi et al., 2013]:

(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 $\boldsymbol{N}_n = (N_1, \ldots, N_{K_n})$
$\iff$ Dirichlet process;

(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 $\boldsymbol{N}_n = (N_1, \ldots, N_{K_n})$
$\iff$ Gibbs–type priors [Gnedin & Pitman, 2006];

(c) $\mathbb{P}[X_{n+1} = \text{“new”} \mid X^{(n)}] = f(n, K_n, \boldsymbol{N}_n, \text{model parameters})$
$\iff$ depends on all information conveyed by the sample i.e. n, K_n and $\boldsymbol{N}_n = (N_1, \ldots, N_{K_n})$
$\iff$ serious tractability issues.

$\Longrightarrow$ *Let's go for the intermediate case which exhibits a richer predictive structure than the DP!*

Gibbs–type priors

Q is a Gibbs-type prior of order $\sigma \in (-\infty, 1)$ if and only if it gives rise to predictive distributions of the form

$$\mathbb{P}[X_{n+1} \in \cdot \mid X^{(n)}]$$

$$= \underbrace{\frac{V_{n+1,K_n+1}}{V_{n,K_n}}}_{\mathbb{P}[X_{n+1}=\text{“new”} \mid X^{(n)}]} \underbrace{P^*(\cdot)}_{\text{prior guess}} + \underbrace{\left(1 - \frac{V_{n+1,K_n+1}}{V_{n,K_n}}\right)}_{\mathbb{P}[X_{n+1}=\text{“old”} \mid X^{(n)}]} \underbrace{\frac{\sum_{i=1}^{K_n}(N_i - \sigma)\,\delta_{X_i^*}(\cdot)}{n - \sigma K_n}}_{\text{weighted emprirical measure}}$$

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

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

$\Longrightarrow$ A Gibbs–type prior is completely characterized by choice of P^*, $\sigma < 1$ and a set of weights $V_{n,j}$'s.

$\Longrightarrow$ Crucially now $\mathbb{P}[X_{n+1} = \text{“new”} \mid X^{(n)}]$ depends on both the sample size n and the distinct values in the sample K_n.

Predictive characterization of the Pitman–Yor process

With $\sigma \in [0,1)$ and $\theta > -\sigma$ or $\sigma < 0$ and $\theta = r|\sigma|$ with $r \in \mathbb{N}$ and

$$V_{n,j} = \frac{\prod_{i=1}^{k-1}(\theta + i\sigma)}{(\theta+1)_{n-1}}$$

one obtains

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

which correspond to the Pitman–Yor (PY) process aka two parameter Poisson–Dirichlet process (Pitman & Yor, 1997)

$\Longrightarrow$ $\mathbb{P}[X_{n+1} = \text{“new”} \mid X^{(n)}]$ is monotonically increasing or decreasing in K_n according to $\sigma > 0$ or $\sigma < 0$, respectively.

If $\sigma = 0$, the PY reduces to the Dirichlet process and

$$\mathbb{P}\left[X_{n+1} \in A \,\middle|\, X^{(n)}\right] = \underbrace{\frac{\theta}{\theta+n}}_{=\mathbb{P}[X_{n+1}=\text{“new”} \mid X^{(n)}]} P^*(A) + \frac{1}{\theta+n}\sum_{i=1}^{K_n} N_i\delta_{X_i^*}(A).$$

A closer look at the Gibbs–type predictive structure

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

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

$$\frac{V_{n+1,K_n+1}}{V_{n,K_n}} \left[\overset{\text{PY case}}{=} \frac{\theta + \sigma K_n}{\theta + n} \right]$$

or "old" $\{X_1^*, \ldots, X_{K_n}^*\}$ with probability

$$1 - \frac{V_{n+1,K_n+1}}{V_{n,K_n}} = \frac{(n - \sigma K_n) V_{n+1,K_n}}{V_{n,K_n}} \left[\overset{\text{PY case}}{=} \frac{n - \sigma K_n}{\theta + n} \right]$$

. $\Longrightarrow$ depends on n and K_n but not on the frequencies $(N_1, \ldots, N_{K_n})$

(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

$$\frac{N_i - \sigma}{n - \sigma K_n}.$$

$\Longrightarrow$ A reinforcement mechanism driven by σ takes place among the "old" values. For instance, if $N_1 = 2$ and $N_2 = 1$ then

$$\frac{\mathbb{P}[X_{n+1} = X_1^* \mid X^{(n)}]}{\mathbb{P}[X_{n+1} = X_2^* \mid X^{(n)}]} = \frac{2 - \sigma}{1 - \sigma} = \begin{cases} < 2 & \text{if } \sigma < 0 \\ 2 & \text{if } \sigma = 0 \\ > 2 & \text{if } \sigma > 0 \end{cases} \overset{e.g.}{=} \begin{cases} 1.5 & \text{if } \sigma = -1 \\ 2 & \text{if } \sigma = 0 \\ 3 & \text{if } \sigma = 0.5 \end{cases}$$

Who are the members of the class of Gibbs–type priors?

Characterization according to the value of σ (Gnedin and Pitman, 2006):

- $\sigma = 0 \iff$ Dirichlet process or Dirichlet process mixed over its total mass parameter $\theta > 0$; recall that $K_n/\log n \xrightarrow{\text{a.s.}} \theta$
- $0 < \sigma < 1 \iff$ 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. Crucially

$$\frac{K_n}{n^\sigma} \xrightarrow{\text{a.s.}} S_\sigma \qquad \text{as } n \to \infty$$

  $\implies$ by tuning σ whole spectrum of growth rates
- $\sigma < 0 \implies$ mixtures of symmetric k–variate Dirichlet distributions

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

  and $K_n \xrightarrow{\text{a.s.}} R_\sigma$.

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.

Special cases

- $\sigma > 0$: In addition to the PY 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 $\Gamma(x,a)$ is the incomplete gamma function. If $\sigma = 1/2$ it reduces to the normalized inverse Gaussian process (N–IG).

- $\sigma < 0$:
  - If π is degenerate on $r \in \mathbb{N}$, one has symmetric r–variate Dirichlet distribution which corresponds to a PY 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 $\gamma \in (0,1)$,

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

    $\Longrightarrow$ Number of species is finite (a.s.) but with infinite mean!
  - Other interesting cases arise if π is a Poisson distribution (restricted to the positive integers) or a geometric distribution.

Induced distribution on number of clusters

An equivalent definition of Gibbs–type priors (Gnedin & Pitman, 2006) is as species sampling models which induce a random partition of the form characterized by

$$\Pi_j^n(n_1,\ldots,n_j) = V_{n,j}\prod_{i=1}^{j}(1-\sigma)_{n_i-1}\left[\overset{\text{PY case}}{=}\frac{\prod_{i=1}^{j-1}(\theta+i\sigma)}{(\theta+1)_{n-1}}\prod_{i=1}^{j}(1-\sigma)_{n_i-1}\right] \quad (\triangle)$$

for any $n \geq 1$, $j \leq n$ and positive integers $n_1,\ldots,n_j$ such that $\sum_{i=1}^{j} n_i = n$.

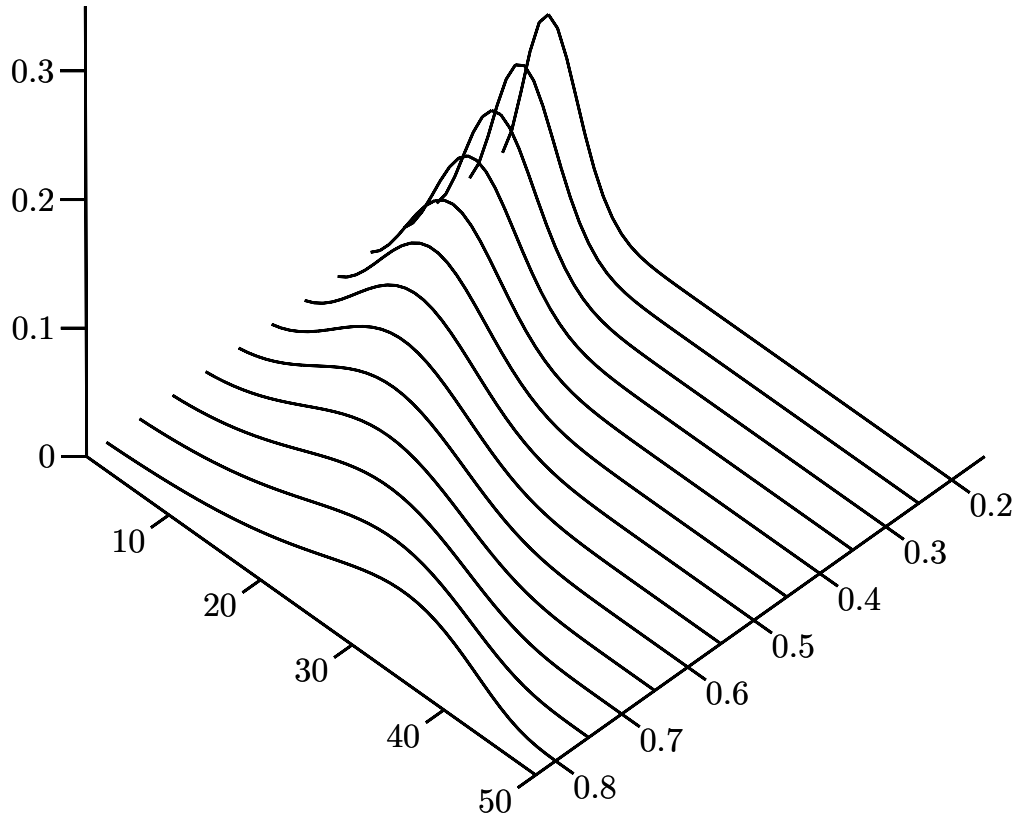
Intepretation of $(\triangle)$: probability of observing a specific sample $X_1,\ldots,X_n$ featuring j distinct observations with frequencies $n_1,\ldots,n_j \implies$ exchangeable partition probability function (EPPF).

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

$$\mathbb{P}(K_n = j) = \frac{V_{n,j}}{\sigma^j}\mathscr{C}(n,j;\sigma)\left[\overset{\text{PY case}}{=}\frac{\prod_{i=1}^{k-1}(\theta+i\sigma)}{(\theta+1)_{n-1}\,\sigma^j}\mathscr{C}(n,j;\sigma)\right]$$

with $\mathscr{C}(n,j;\sigma) = \frac{1}{j!}\sum_{i=0}^{j}(-1)^i\binom{j}{i}(-i\sigma)_n$ denoting a generalized factorial coefficient.

Prior distribution of the number of clusters as σ varies



Prior distributions on the number of clusters corresponding to the PY process with $n = 50, \theta = 1$ *and* $\sigma = 0.2, 0.3, \ldots, 0.8$.

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 PY and β in the NGG case) controls the location of the (prior) distribution of K_n

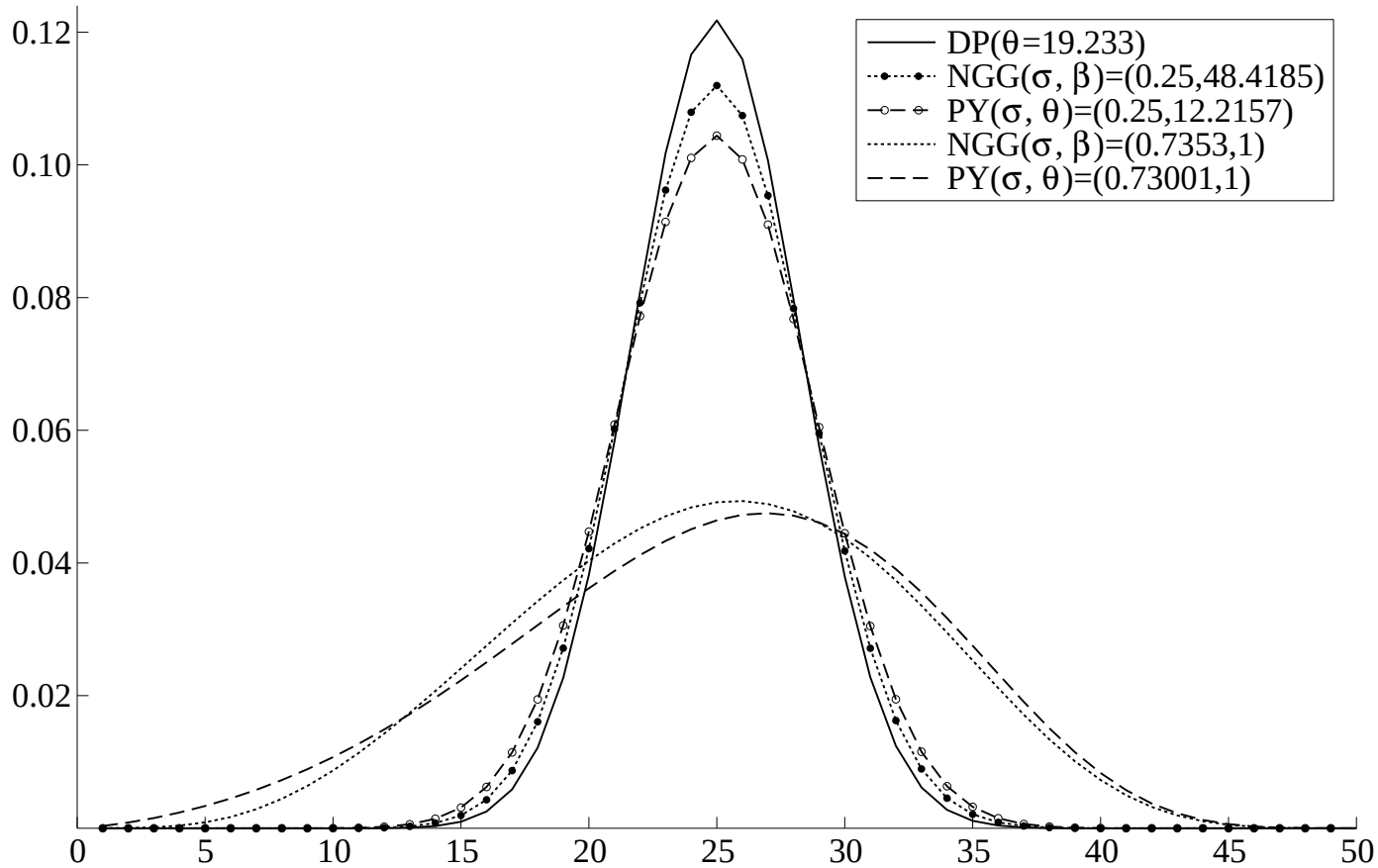
Example of distributions of K_n induced by different Gibbs–type priors:

- $n = 50$ and the prior expected number of clusters is 25 $\implies$ fix the prior parameters s.t. $\mathbb{E}(K_{50}) = 25$.
- 5 different models:
  - Dirichlet process with $\theta = 19.23$;
  - PY processes with $(\sigma, \theta) = (0.73, 1)$ and $(\sigma, \theta) = (0.25, 12.22)$;
  - NGG processes with $(\sigma, \beta) = (0.74, 1)$ and $(0.25, 48.42)$.

$\implies$ 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 PY and the NGG processes. The values of the parameters are set in such a way that $\mathbb{E}(K_{50}) = 25$.

Toy mixture example

A popular use of discrete $\tilde{P}$ is within hierarchical mixture models for density estimation and clustering. The latter is carried out at the latent level and makes use of the discrete nature of $\tilde{P}$.

- $n = 50$ 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

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

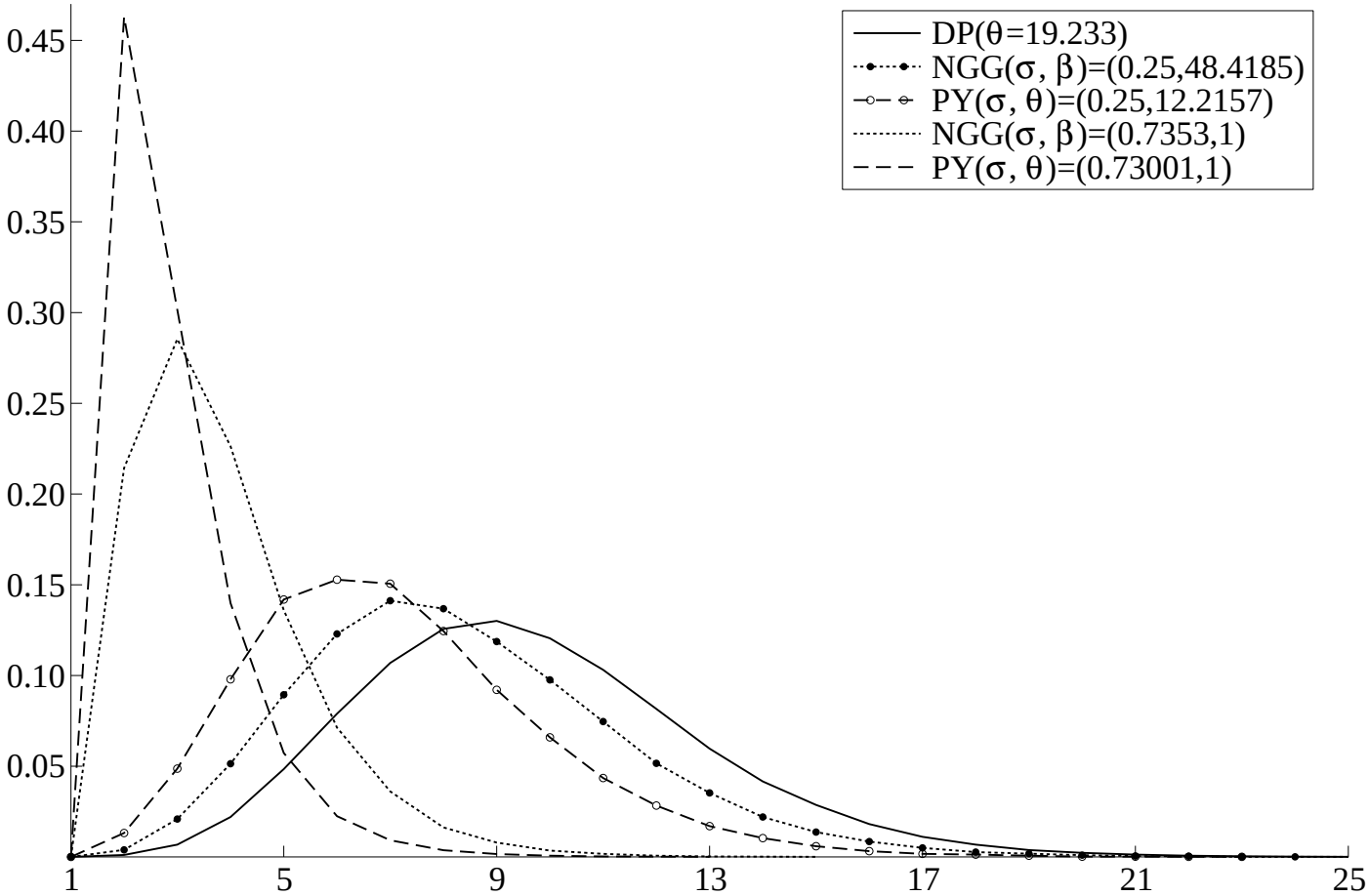
  with Q a discrete nonparametric prior.
- The distribution of K_n represents the prior distribution on the number of mixture components; some summary statistics of the posterior $(K_n | Y^{(n)})$ is then used as estimate of the number of mixture components.

- We select priors Q with "misspecified" parameters: in particular, the ones of the previous example which imply $\mathbb{E}[K_{50}] = 25$ i.e. a prior opinion on K_{50} remarkably far from the true number of components, namely 2.
- Recall the 5 different models:
  - Dirichlet process with $\theta = 19.23$;
  - PD processes with $(\sigma, \theta) = (0.73, 1)$ & $(\sigma, \theta) = (0.25, 12.22)$;
  - NGG processes with $(\sigma, \beta) = (0.74, 1)$ & $(0.25, 48.42)$.

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

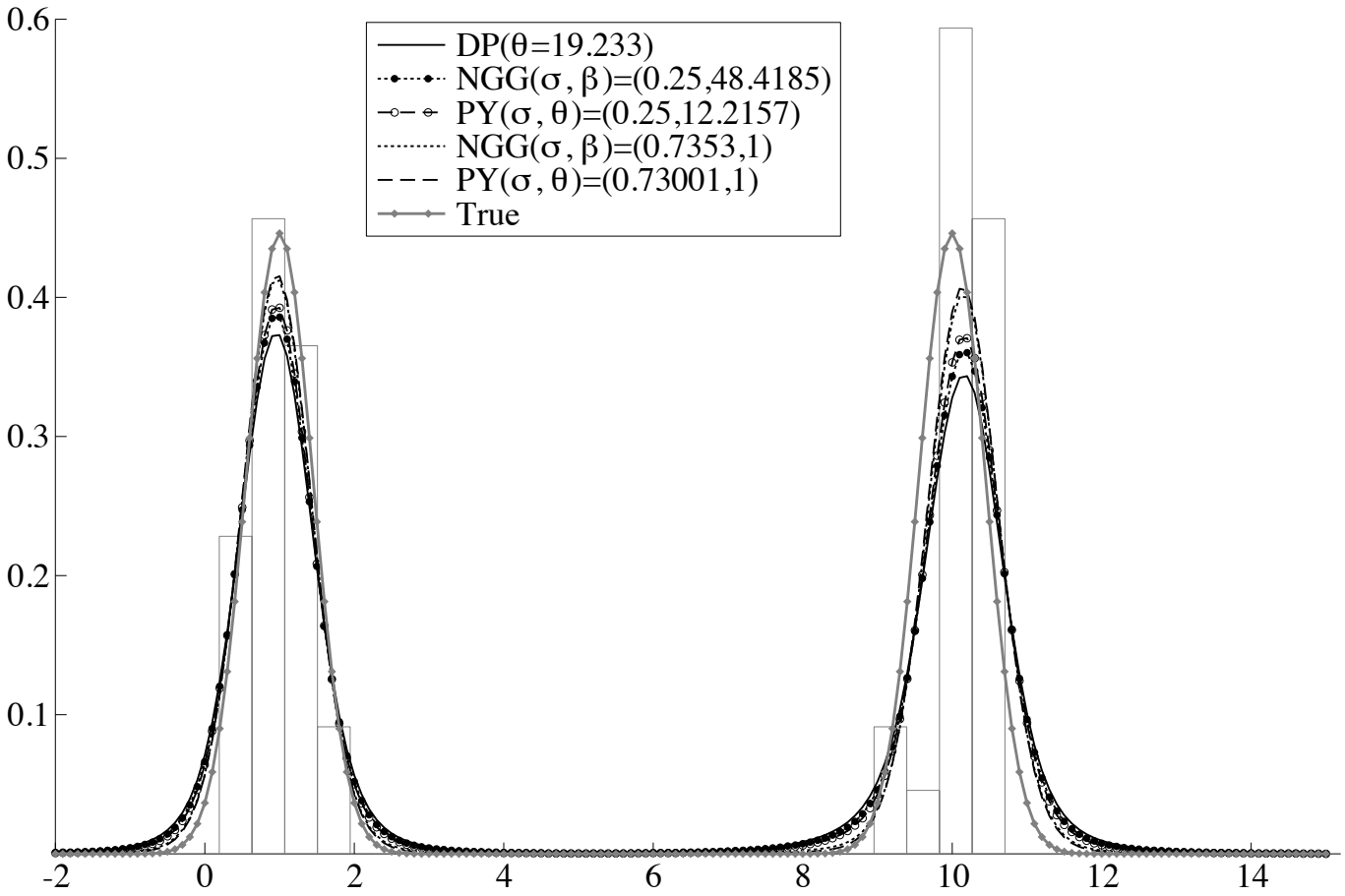
$\Longrightarrow$ the larger σ the better is the posterior estimate of K_n.
$\Longrightarrow$ in terms of density estimation the difference is negligible; this is because one can always fit a mixture density with more components than needed.

Posterior distribution of the number of mixture components



Posterior distributions on the number of groups corresponding to 5 mixture models.

Posterior density estimates



Density estimates corresponding to the 5 mixture models.

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 $\boldsymbol{N} = (N_1, \ldots, N_{K_n})$ s.t. $\sum_{i=1}^{K_n} N_i = n$;
  - ◇ the labels (names) X_i^*'s of the distinct species, for $i = 1, \ldots, K_n$.

- The information provided by $\boldsymbol{N}$ can also be coded by $\boldsymbol{M} := (M_1, \ldots, M_n)$

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

  Note that $\sum_{i=1}^{n} M_i = K_n$ and $\sum_{i=1}^{n} iM_i = n$.

- 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

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

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

One–step Prediction

- Discovery probability estimation: Given a basic sample $X^{(n)}$, estimate the probability of discovering at the (n+1)–th sampling step either a new species or an "old" species with frequency r.
- Turing estimator [Good, 1953; Mao & Lindsay, 2002]: probability of discovering at (n+1)–th step a new species is

$$\frac{m_1}{n}$$

and a species with frequency r in $X^{(n)}$ is

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

- Problem: m_{r+1} is used to estimate the probability of discovering a species with frequency $r \Longrightarrow$ counterintuitive! It should be based on m_r.
  E.g. If $m_5 = 10, m_6 = 0$, the estimated probability of detecting a species with frequency 5 would be $0 \Longrightarrow$ problem bypassed by use of "smoothing functions" but, in I.J. Good's words, it seems like an "adhockery"!
  Origin of the problem? In a frequentist nonparametric setup there is no natural quantity to use for estimating the probability of discovering a new species and so m_1 is used. Hence, the discovery probability of species with frequency 1 uses m_2 and so on.

BNP approach to discovery probab. [Lijoi, Mena & P, 2007a]

> *Key advantage of the Bayesian nonparametric approach is that the predictive structure includes a positive probability of discovering a new species, value, cluster, agent etc.*

Assume the data $(X_n)_{n\geq 1}$ are exchangeable with a Gibbs–type de Finetti measure.

BNP analog to Turing estimator: given a basic sample $X^{(n)}$ featuring $K_n = j$ distinct species with $m_1, \ldots, m_n$ s.t. $\sum_{i=1}^{n} i\, m_i = n$:

- the probability of discovering a new species is

$$\mathbb{P}[X_{n+1} = \text{“new”} \mid X^{(n)}] = \frac{V_{n+1,j+1}}{V_{n,j}} \qquad \left[\overset{\text{PY case}}{=} \frac{\theta + \sigma j}{\theta + n} \right].$$

- the probability of detecting a species with frequency r in $X^{(n)}$ is

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

$$\left[\overset{\text{PY case}}{=} \frac{r-\sigma}{\theta+n}\, m_r \right]$$

$\Longrightarrow$ Probability of sampling a species with frequency r depends on m_r!

m–step ahead discovery probability estimation

Conditionally on a basic sample $X^{(n)}$, estimate the probability of discovering at the (n+m+1)–th step either a new species or an "old" species with frequency r without observing the additional sample $X^{(m)} := (X_{n+1}, \ldots, X_{n+m})$.

Remark. From such an estimator one immediately obtains:

- 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 $\implies$ 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 with a sample of size $n + m$.

Frequentist estimators:

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

Unconditional and conditional distributional results for K_n

Restrict attention to the PY process but most results carry over (with minor modifications) to general Gibbs–type priors.

Recall that $K_m^{(n)}$ is the number of new species to be recorded in the additional sample of size m given $X^{(n)}$ featuring $K_n = j$ distinct values.

- If $\sigma < 0$ and $\theta = r|\sigma|$, then $K_n \xrightarrow{\text{a.s.}} r$ as $n \to \infty$.
  Also, conditionally on $X^{(n)}$ featuring $K_n = j$ distinct values,

$$K_m^{(n)} | X^{(n)} \xrightarrow{\text{a.s.}} r - j \qquad \text{for} m \to \infty.$$

- If $\sigma = 0$ recall that, as $n, m \to \infty$,

$$\frac{K_n}{\log n} \xrightarrow{\text{a.s.}} \theta \qquad \text{and} \qquad \frac{K_m^{(n)}}{\log m} \,\Big|\, X^{(n)} \xrightarrow{\text{a.s.}} \theta.$$

- If $\sigma > 0$

$$\frac{K_n}{n^\sigma} \xrightarrow{\text{a.s.}} Y_{\theta/\sigma} \qquad (n \to \infty).$$

  where Y_q, with $q \geq 0$ is a generalized Mittag–Leffler random variable.

Still in the case of $\sigma > 0$ and conditionally on $X^{(n)}$, one also has

$$\mathbb{P}\left[K_m^{(n)} = k \middle| X^{(n)}\right] = \frac{(\theta+1)_{n-1}}{(\theta+1)_{n+m-1}} \frac{\prod_{i=j}^{j+k-1}(\theta+i\sigma)}{\sigma^k} \mathscr{C}(m,k;\sigma,-n+j\sigma).$$

Moreover, the expected number of new species in an additional sample is

$$\mathbb{E}[K_m^{(n)}|X^{(n)}] = \left(j+\frac{\theta}{\sigma}\right)\left\{\frac{(\theta+n+\sigma)_m}{(\theta+n)_m} - 1\right\}.$$

See Lijoi, Mena & P (2007a) and Favaro, Lijoi, Mena & P (2009).

As the size of the additional sample m diverges

$$\frac{K_m^{(n)}}{m^\sigma} \middle| X^{(n)} \xrightarrow{\text{a.s.}} Z_{n,j} \qquad (m \to \infty)$$

where $Z_{n,j} \stackrel{d}{=} B_{j+\theta/\sigma,\, n/\sigma-j}\, Y_{(\theta+n)/\sigma}$ with $B_{a,b}$ a beta random variable independent of Y_q. [Favaro, Lijoi, Mena & P, 2009]

$\Longrightarrow$ *In the DP case logarithmic growth of K_n: such a restriction has been overcome by allowing a richer predictive structure and the growth now depends on the model parameter σ.*

m–step ahead BNP estimators based on the PY process

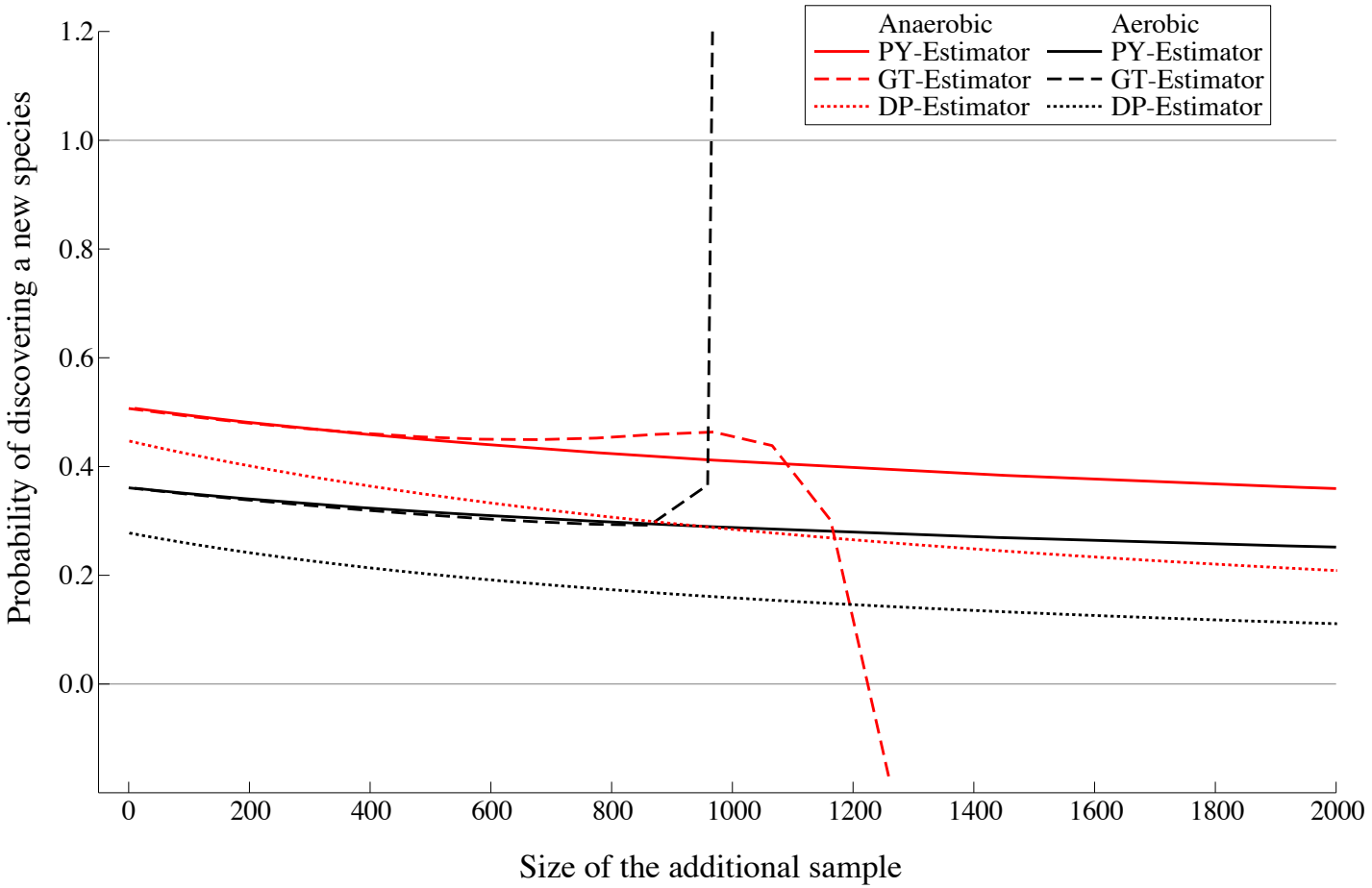
- BNP analog of the Good–Toulmin estimator [Favaro, Lijoi, Mena & P, 2009]: estimator for the probability of discovering a new species at the (n+m+1)–th step

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

- BNP estimator for the probability of discovering a species with frequency r at the (n+m+1)–th sampling step [Favaro, Lijoi & P, 2012a]:

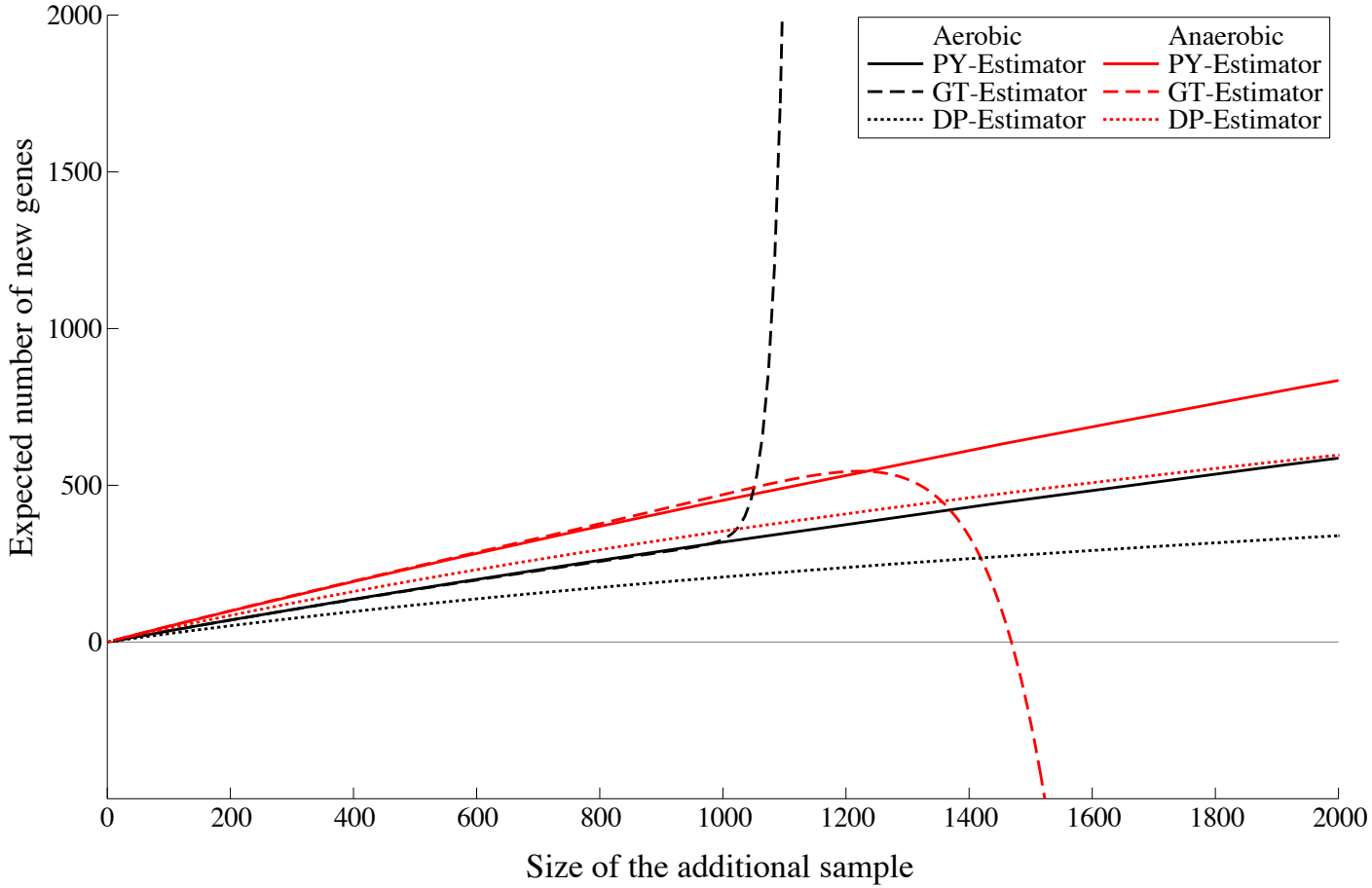
$$\begin{aligned}\mathbb{P}[X_{n+m+1} = \text{“species with frequency } r\text{”} \mid X^{(n)}] \\ = \sum_{i=1}^{r} m_i (i-\sigma)_{r+1-i} \binom{m}{r-i} \frac{(\theta + n - i + \sigma)_{m-r+i}}{(\theta + n)_{m+1}} \\ + (1-\sigma)_r \binom{m}{r} \frac{(\theta + k\sigma)(\theta + n + \sigma)_{m-r}}{(\theta + n)_{m+1}}\end{aligned}$$

Discovery probability at the (n+m+1)–th sampling step



EST data from Naegleria gruberi aerobic and anaerobic cDNA libraries with basic sample $n \cong 950$*: Good–Toulmin (GT), DP process and PY process 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 additional sample of size m.



EST data from Naegleria gruberi aerobic and anaerobic cDNA libraries with basic sample $n \cong 950$*: Good–Toulmin (GT), DP process and PY process 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 based on Gibbs–type priors available for the before mentioned and other quantities of interest in species sampling problems.
- 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 $\sigma < 0$ (surprising heuristic by–product: by combining Gibbs-type priors with $\sigma > 0$ and $\sigma < 0$ is possible to identify situations in which frequentist estimators work).

Full weak support property of Gibbs–type priors

Henceforth focus on "genuinely nonparametric priors": $\sigma \geq 0$ or $\sigma < 0$ with the distribution π on the number of blocks of the symmetric Dirichlet distribution having support $\mathbb{N}$

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

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

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

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?

Definition. *Q is weakly consistent at P_0 if for every $\varepsilon > 0$*

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

where A_ε is a weak neighbourhood of P_0 and P_0^∞ denotes the infinite product measure.

This setup essentially consists in postulating that Q is our model and we update it *as if* the data $(X_n)_{n\geq 1}$ were an exchangeable sequence generated by Q. But the data are actually i.i.d. from a "true" P_0. This implies that:

- if P_0 is diffuse, all data are distinct and $\kappa_n = n$;
- if P_0 is discrete, there are ties in the data and $\kappa_n = o(n)$.

Notation: To highlight that K_n is dictated by P_0 and not anymore by Q, above and in the following it is denoted by κ_n.

In studying frequentist posterior consistency, but actually all properties of Gibbs–type priors, the key quantity to look at is the probability of discovering a new species

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

Recall that in the Gibbs-framework it depends n and κ_n but not on $\boldsymbol{N}_n = (N_1, \ldots, N_{\kappa_n})$. For instance, in the two parameter PD case it is

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

Proof strategy for frequentist consistency consists in showing that

- $\mathbb{E}[\tilde{P} \mid X^{(n)}] \overset{n\to\infty}{\longrightarrow} P_0$ a.s.–P_0^∞ $\iff$ by the predictive structure of Gibbs–type priors this is equivalent to showing that

$$\mathbb{P}[X_{n+1} = \text{“new”} \mid X^{(n)}] = V_{n+1,k+1}/V_{n,k} \overset{n\to\infty}{\longrightarrow} 0 \quad \text{a.s.–}P_0^\infty$$

- $\text{Var}[\tilde{P} \mid X^{(n)}] \overset{n\to\infty}{\longrightarrow} 0$ a.s.–P_0^∞ by finding a suitable bound on the variance.

We first investigate consistency for Gibbs–type priors with $\sigma \in (-\infty, 0)$

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 $x \in \mathbb{X}$)

Let Q be a Gibbs–type prior with $\sigma < 0$, mixing measure π and P_0 a discrete "true" distribution. If for sufficiently large x

$$\frac{\pi(x+1)}{\pi(x)} \leq 1, \qquad (\blacktriangledown)$$

then Q is consistent at P_0.

Remark. ($\blacktriangledown$) serves only for pinning down the proof in general.

$\Longrightarrow$ 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 diffuse "true" data generating distribution P_0

Diffuse $P_0 \implies$ wide range of asymptotic behaviours including erratic ones.

Remark. Recall that, since P_0 is diffuse, $\kappa_n = n$.

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:

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

$$= \frac{\lambda n}{(2n+1)(2n)} \frac{{}_1F_1(n; 2n; \lambda)}{{}_1F_1(n+1; 2n+2; \lambda)} \sim \frac{\lambda}{2(2n+1)} \overset{n\to\infty}{\longrightarrow} 0$$

This, combined with some other arguments, shows that such a prior is consistent at any diffuse P_0.
Example 2: Gnedin's model with $\sigma = -1$ and parameter $\gamma \in (0, 1)$.
For diffuse P_0 we obtain:

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

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

Example 3: Gibbs–type prior with $\sigma = -1$ and geometric(η) mixing dist. π. For diffuse 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 diffuse P_0 is then as follows:

> *Let Q be a Gibbs–type prior with $\sigma < 0$ and P_0 a diffuse "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 case of $\sigma > 0$

For the case of $\sigma > 0$ one has the following partial result:

- *General case: Let Q be a Gibbs–type prior with $\sigma > 0$, prior guess $P^* = \mathbb{E}[\tilde{P}]$ and P_0 be the "true" distribution. Assume that $\mathbb{P}(X_{n+1} = \text{"new"} \mid X^{(n)}) = \frac{V_{n+1,\kappa_n+1}}{V_{n,\kappa_n}}$ converges a.s.-P_0^∞. Then $Q(\cdot \mid X^{(n)})$ converges weakly, a.s.-P_0^∞, to a point mass at*

$$\alpha P^*(\cdot) + (1-\alpha)P_0(\cdot) \text{with} \qquad \alpha \in [0,1].$$

- *Q is a PY process or a normalized generalized gamma process prior (James, 2008, Jang, Lee and Lee, 2010): consistency holds if P_0 is discrete, otherwise one has inconsistency.*

Remark: Recall that in two parameter PD case we have

$$\mathbb{P}[X_{n+1} = \text{"new"} \mid X^{(n)}] = \frac{\theta + \sigma\kappa_n}{\theta + n} = \begin{cases} \to 0 & \text{if } \kappa_n = o(n) \\ \to \sigma & \text{if } \kappa_n = n \end{cases}$$

In the Dirichlet case $\sigma = 0$ and, whatever the behaviour of κ_n, one has

$$\mathbb{P}[X_{n+1} = \text{"new"} \mid X^{(n)}] = \frac{\theta}{\theta + n} \to 0$$

$\Longrightarrow$ consistent for both discrete and diffuse "true" distributions.

What does this asymptotic analysis tell us?

Discrete $\tilde{P}$ designed to model discrete distributions and should not be used to model data from diffuse distributions.

Remark. Dirichlet process enjoys:
◇ full weak support property
◇ weak consistency for diffuse 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 diffuse 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!

$\implies$ There are not good or bad priors, but rather models which are suitable/compatible/designed for a certain phenomenon and not for others!

Some References

• Antoniak(1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Statist.* **2**, 1152–1174.

• 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.

• De Blasi, Lijoi, & Prünster (2013). An asymptotic analysis of a class of discrete nonparametric priors. *Satist. Sinica* **23**, 1299–1322.

• Doksum (1974). Tailfree and neutral random probabilities and their posterior distributions. *Ann. Probab.* **2**, 183–201.

• Caron (2012). Bayesian nonparametric models for bipartite graphs. NIPS'2012.

• Caron & Fox (2017), Sparse graphs using exchangeable random measures. *J. R. Stat. Soc. B* **79**, 1295-1366.

• Diaconis & Freedman (1986). On the consistency of Bayes estimates. *Ann. Statist.* **14**, 1-26.

• Favaro, Lijoi, Mena & Prünster (2009). Bayesian nonparametric inference for species variety with a two parameter Poisson-Dirichlet process prior. *J. Roy. Statist. Soc. Ser. B* **71**, 993–1008.

• Favaro, Lijoi & Prünster (2012a). A new estimator of the discovery probability. *Biometrics* **68**, 1188-96.

• Favaro, Lijoi & Prünster (2012b). On the stick–breaking representation of normalized inverse Gaussian priors. *Biometrika* **99**, 663-674.

• Favaro, Prünster & Walker (2011). On a class of random probability measures with general predictive structure. *Scand. J. Statist.* **38**, 359–376.

• Ferguson (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1**, 209-30.
• Ferguson (1974). Prior distributions on spaces of prob. meas. *Ann. Stat.* **2**, 615-29.
• Fortini, Ladelli, Regazzini (2000). Exchangeability, predictive distributions and parametric models. Sankhya A **62**, 86-109.
• 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**, 5674-85.
• 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.
• James (2008). Large sample asymptotics for the two parameter Poisson Dirichlet process. In *Pushing the Limits of Contemporary Statistics* (Clarke & Ghosal eds.), IMS, Hayward, pp. 187-199.
• Jang, Lee & Lee (2010). Posterior consistency of species sampling priors. *Statist. Sinica* **20**, 581-593.
• Korwar & Hollander (1973). Contribution to the theory of Dirichlet processes. *Ann. Probab.* **1**, 705-11.
• Lijoi, Mena & Prünster (2005). Hierarchical mixture modelling with normalized inverse Gaussian priors. *J. Amer. Statist. Assoc.* **100**, 1278-1291.
• Lijoi, Mena & Prünster (2007a). Bayesian nonparametric estimation of the probability of discovering new species. *Biometrika* **94**, 769–786.

• Lijoi, Mena & Prünster (2007b). Controlling the reinforcement in Bayesian nonparametric mixture models. *J. Roy. Statist. Soc. Ser. B* **69**, 715–740.
• Lo (1991). A characterization of the Dirichlet process. *Statist. Probab. Lett.* **12**, 185–187.
• Mao (2004). Prediction of the conditional probability of discovering a new class. *J. Am. Statist. Assoc.* **99**, 1108-18.
• Mao & Lindsay (2002). A Poisson model for the coverage problem with a genomic application. *Biometrika* **89**, 669-81.
• Perman, Pitman & Yor (1992). Size-biased sampling of Poisson point processes and excursions. *Probab. Theory Related Fields* **92**, 21-39.
• Pitman (1995).Exchangeable and partially exchangeable random partitions. *Prob. Th. and Rel. Fields* **102**, 145–158.
• Pitman & Yor (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25**, 855–900.
• Pitman (2006). *Combinatorial Stochastic Processes*. Lecture Notes in Math., vol.1875, Springer, Berlin.
• Regazzini (1978). Intorno ad alcune questioni relative alla definizione del premio secondo la teoria della credibilità. *Giorn. Istit. Ital. Attuari*, **41**, 77–89.
• Regazzini & Petris (1992). Some critical aspects of the use of exchangeability in statistics. J. Ital. Statist. Soc., **1**, 103–130.
• Sethuraman (1994). A constructive definition of the DP prior. *Stat. Sin.* **2**, 639–50.
• Teh (2006). A Hierarchical Bayesian Language Model based on Pitman-Yor Processes. *Coling/ACL 2006*, 985-92.