Two tales about Bayesian nonparametric modeling

Igor Prünster

University of Torino & Collegio Carlo Alberto

NIPS Workshops

Bayesian Nonparametric Methods: Hope or Hype?

Sierra Nevada, 17th December 2011

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

Outline

The Bayesian nonparametric framework

Induction & Exchangeability: Prediction under symmetry (i.e. homogeneity, analogy) between past and future

$\Longleftrightarrow$ assumption of exchangeability, i.e. $(X_n)_{n\geq 1}$ is exchangeable if for every finite permutation π

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

.

The Bayesian nonparametric framework

Induction & Exchangeability: Prediction under symmetry (i.e. homogeneity, analogy) between past and future

$\Longleftrightarrow$ assumption of exchangeability, i.e. $(X_n)_{n\geq 1}$ is exchangeable if for every finite permutation π

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

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

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

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 $\tilde{P}$.

Equivalently one can state that

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

Depending on the structure of 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 (Ferguson, 1974): Q should have "large" support (possibly the whole $\mathscr{P}$)

Depending on the structure of 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 (Ferguson, 1974): Q should have "large" support (possibly the whole $\mathscr{P}$)

(One–step) Prediction is then achieved by

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

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

Discrete nonparametric priors

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

$$\tilde{P}(\,\cdot\,) = \sum_{i\geq} \tilde{p}_i \delta_{Y_i}(\,\cdot\,), \qquad (*)$$

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 (RPM)

$$\tilde{P}(\,\cdot\,) = \sum_{i\geq} \tilde{p}_i \delta_{Y_i}(\,\cdot\,), \qquad (*)$$

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

Discrete nonparametric priors

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

$$\tilde{P}(\,\cdot\,) = \sum_{i\geq} \tilde{p}_i \delta_{Y_i}(\,\cdot\,), \qquad (*)$$

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. 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 when dealing with discrete RPMs is the probability of discovering a new species

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

Probability of discovering a new species

A key quantity when dealing with discrete RPMs is the probability of discovering a new species

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

Discrete RPMs can be classified in the 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 $\boldsymbol{N}_n = (N_1, \ldots, N_{K_n})$

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

Probability of discovering a new species

A key quantity when dealing with discrete RPMs is the probability of discovering a new species

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

Discrete RPMs can be classified in the 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 $\boldsymbol{N}_n = (N_1, \ldots, N_{K_n})$
$\iff$ 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 $\boldsymbol{N}_n = (N_1, \ldots, N_{K_n})$
$\iff$ Gibbs–type priors (Gnedin and Pitman, 2006);

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

Complete predictive structure

The predictive distributions associated with the Dirichlet process coincide with

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

with "prior guess at the shape" $P^* := \mathbb{E}(\tilde{P})$.

Complete predictive structure

The predictive distributions associated with the Dirichlet process coincide with

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

with "prior guess at the shape" $P^* := \mathbb{E}(\tilde{P})$.

The two parameter Poisson–Dirichlet process or Pitman–Yor process (Perman, Pitman & Yor, 1992) yields predictive distributions of the form

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

with either $\sigma \in [0,1)$ and $\theta > -\sigma$ or $\sigma \in (-\infty, 0)$ and $\theta = r|\sigma|$ with $r \in \mathbb{N}$.
$\Longrightarrow$ if $\sigma = 0$, the Pitman–Yor process reduces to the Dirichlet process.

Gibbs–type random probability measures (Gnedin & Pitman '06)

The Dirichlet process and the two parameter PD process both belong to the general family of Gibbs–type random probability measures.

$\tilde{P}$ 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

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

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

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

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 $\boldsymbol{N}_n = (N_1,\ldots,N_{K_n})$.

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" $\{X_1^*,\ldots,X_{K_n}^*\}$ 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 $\boldsymbol{N}_n = (N_1,\ldots,N_{K_n})$.

(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

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

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 $\Longrightarrow$ "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 $\Longrightarrow$ "genuinely nonparametric priors"

Theorem 1. *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*

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

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

$\Updownarrow$

Weak neighborhoods of any given distribution have *a priori* positive probability

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 A_ε*

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

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

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 A_ε*

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

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

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

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

Theorem 2. *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")

Theorem 2. *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).

$\Longrightarrow$ Theorem 2 guarantees frequentist consistency 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 a 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.
This is illustrated by means of 3 examples:

Remark. Since P_0 is continuous, the number of distinct observations in a sample of size n, K_n, is precisely n.

Example 1: Gibbs–type prior with $\sigma = -1$ related to a Poisson(λ) distribution.
Let us look at the key quantity given by 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 continuous P_0.

Example 2: Gnedin's (2010) Gibbs–type prior with parameter $\gamma \in (0, 1)$.
For continuous P_0 we obtain:

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

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 (2010) Gibbs–type prior with parameter $\gamma \in (0,1)$.
For continuous P_0 we obtain:

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

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 $\sigma = -1$ related to a geometric(η) distribution.
For continuous P_0 we obtain:

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

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

$\Longrightarrow$ 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 ($\eta = 1$).

Gnedin & Pitman (2006) showed that a Gibbs–type prior with $\sigma < 0$ can be represented as mixture of symmetric Dirichlet distributions i.e.

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

Using the species metaphor, this model corresponds to putting a prior π on the number of species K and, conditionally on the number of species being $K = x$, these are distributed as a x–variate symmetric Dirichlet distribution.

Remark. For $\sigma \geq 0$ the model assumes the existence of an infinite number of species, whereas for $\sigma < 0$ it assumes a random but finite number of species. E.g. Gnedin's model: number of species a.s. finite with infinite mean.

Gnedin & Pitman (2006) showed that a Gibbs–type prior with $\sigma < 0$ can be represented as mixture of symmetric Dirichlet distributions i.e.

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

Using the species metaphor, this model corresponds to putting a prior π on the number of species K and, conditionally on the number of species being $K = x$, these are distributed as a x–variate symmetric Dirichlet distribution.

Remark. For $\sigma \geq 0$ the model assumes the existence of an infinite number of species, whereas for $\sigma < 0$ it assumes a random but finite number of species. E.g. Gnedin's model: number of species a.s. finite with infinite mean.

The general consistency results for continuous P_0 is then as follows:
Theorem 3. *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 $M < \infty$*

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

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

Gnedin & Pitman (2006) showed that a Gibbs–type prior with $\sigma < 0$ can be represented as mixture of symmetric Dirichlet distributions i.e.

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

Using the species metaphor, this model corresponds to putting a prior π on the number of species K and, conditionally on the number of species being $K = x$, these are distributed as a x–variate symmetric Dirichlet distribution.

Remark. For $\sigma \geq 0$ the model assumes the existence of an infinite number of species, whereas for $\sigma < 0$ it assumes a random but finite number of species. E.g. Gnedin's model: number of species a.s. finite with infinite mean.

The general consistency results for continuous P_0 is then as follows:
Theorem 3. *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 $M < \infty$*

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

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

Remark. The "extremely mild" technical condition of Theorem 2 corresponds to asking π to be ultimately decreasing.

Implications for Bayes Nonparametrics

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.

Implications for Bayes Nonparametrics

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 RPM designed to model discrete distrib. and should not be used to model data from continuous distributions.

Implications for Bayes Nonparametrics

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 RPM 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
  $\Longrightarrow$ discrete RPMs OK also for continuous distributions: misleading!
  $\Longrightarrow$ inconsistency example in Diaconis & Freedman (1986) interpreted as need to be careful with BNP in general: misunderstanding!

  Why misunderstanding? As the sample size n diverges:
  ◇ P_0 generates $(X_n)_{n\geq 1}$ containing no ties with probability 1
  ◇ a discrete RPM generates $(X_n)_{n\geq 1}$ containing no ties with probability 0
  $\Longrightarrow$ model and data generating mechanism are incompatible!
  Consistency at continuous P_0 is irrelevant: it is just a coincidence if they are (Dirichlet, Gibbs with Poisson mixing).

- Extension to survival analysis: for continuous P_0 both beta-Stacy (Walker & Muliere, 1997) and beta (Hjort, 1990) processes should not be used even if consistent (Kim & Lee, 2001)
  $\Longrightarrow$ use mixtures driven by such processes!
  For discrete P_0 they are fine like all other members of the families of RPM they belong to and which are inconsistent for continuous P_0.

- Extension to survival analysis: for continuous P_0 both beta-Stacy (Walker & Muliere, 1997) and beta (Hjort, 1990) processes should not be used even if consistent (Kim & Lee, 2001)
  $\Longrightarrow$ use mixtures driven by such processes!
  For discrete P_0 they are fine like all other members of the families of RPM they belong to and which are inconsistent for continuous P_0.

- Take–home–message: BNP models are typically consistent in situations they are designed for. Moreover:
  (a) mathematics is tough and so many cases are not covered by the provided sufficient conditions: these are to be interpreted as " most likely " being consistent;
  (b) inconsistent behaviour in BNP is typically solved being " more nonparametric " (e.g. requiring full Kullback–Leibler support instead of full weak support).

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 $i = 1, \ldots, K_n$.

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 $\sum_{i=1}^{n} M_{i,n} = K_n$ 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

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

  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 $X^{(n)}$.

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.

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).

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

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

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 $K_n = k$.

Remark 2. Estimated discovery as r increases:
• Turing estimator produces irregular behaviours (abrupt decays and steady periods) $\Longrightarrow$ based on raw frequencies;
• BNP estimator typically produces a smooth decays $\Longrightarrow$ based on a well–structured underlying probabilistic model.

- 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 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.

- 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.

- BNP estimator for rare species discovery: if species appearing less than τ times are considered rare, then the estimator is given by

$$\mathbb{P}[X_{n+m+1} = \text{“new”} | X^{(n)}] + \sum_{r=1}^{\tau} \mathbb{P}[X_{n+m+1} = \text{species with frequency } r | X^{(n)}].$$

The discovery probability in the Pitman–Yor case

The natural candidate for applications is the Pitman–Yor 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 Pitman–Yor process allows for rates n^σ for $\sigma \in (0, 1)$. See also Teh (2006).

The discovery probability in the Pitman–Yor case

The natural candidate for applications is the Pitman–Yor 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 Pitman–Yor process allows for rates n^σ for $\sigma \in (0,1)$. See also Teh (2006).

- PY 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

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

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

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

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

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

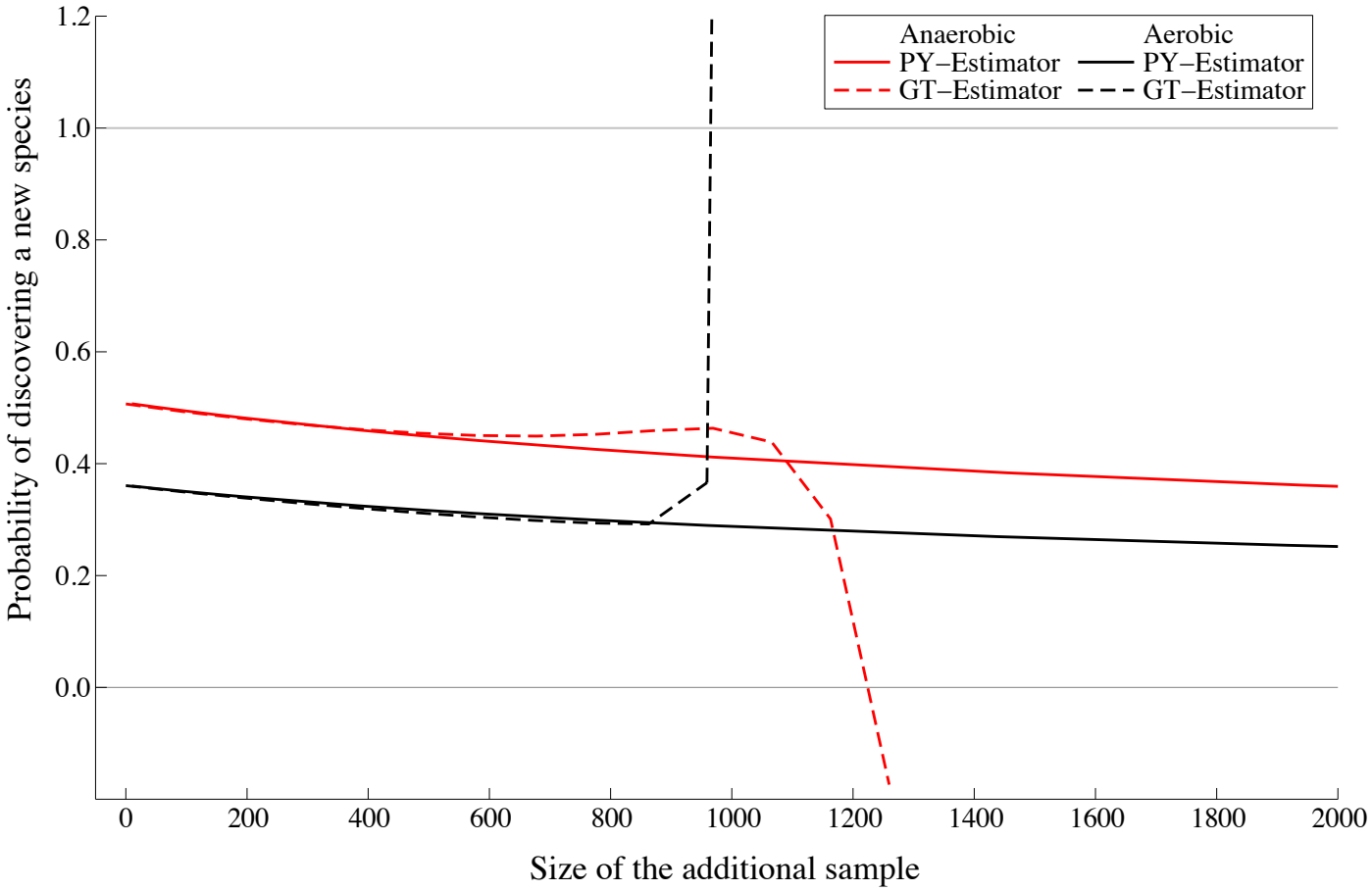
- PY estimator for the probability of discovering a species with frequency r at the (n+m+1)–th step

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

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

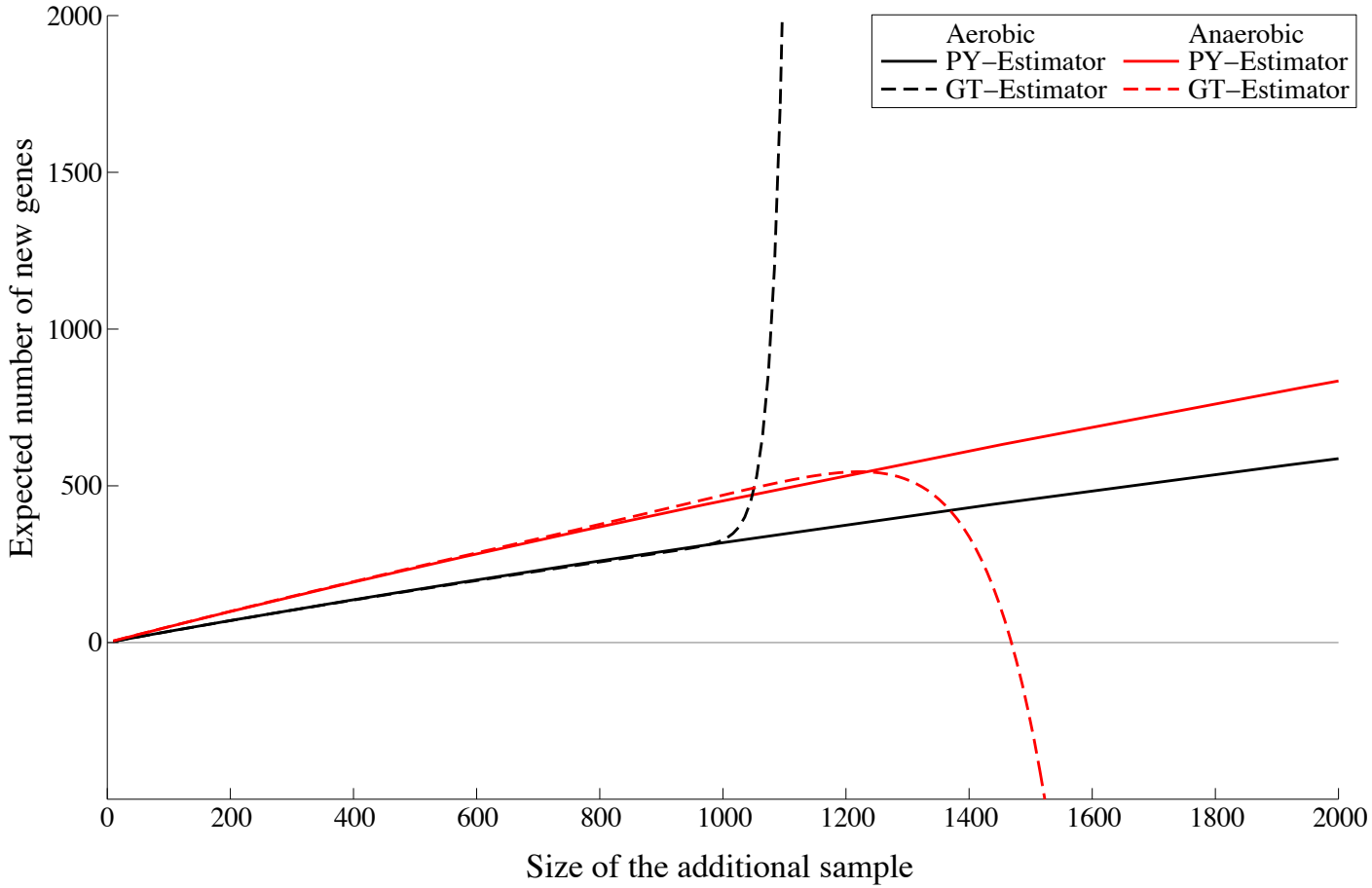
$$+ \frac{(1 - \sigma)_r}{(\theta + n)_{m+1}} \left[(\theta + k\sigma)(\theta + n + \sigma)_{m-r} - \prod_{i=k}^{k+m-r} (\theta + i\sigma) \right]$$

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 Pitman–Yor (PY) 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

- BNP estimators available for other quantities of interest in species sampling problems (completely explicit in the Pitman–Yor 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

- BNP estimators available for other quantities of interest in species sampling problems (completely explicit in the Pitman–Yor 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.
- Pitman–Yor process and Gibbs–type priors with $\sigma > 0$ correspond to models with infinite species (as n diverges also K_n diverges):
  - model is ideally suited for populations with large unknown number of species $\Longrightarrow$ typical case in Genomics;
  - in Ecology "∞" assumption often too strong: problem is not of BNP in general but again related to suitability of the particular BNP model $\Longrightarrow$ Gibbs–type priors with $\sigma < 0$ allow for a large number of species which is finite and random (*work in progress*);
  - Surprising by–product: by combining Gibbs-type priors with $\sigma > 0$ and $\sigma < 0$ is possible to identify situations in which frequentist estimators work.

BNP: Hope or Hype?

My personal answer: BNP is hope but efforts are needed in trying to understand the underlying probabilistic structures and what kind of data generating mechanism each model is designed for.

BNP: Hope or Hype?

My personal answer: BNP is hope but efforts are needed in trying to understand the underlying probabilistic structures and what kind of data generating mechanism each model is designed for.

THANK YOU!

References

• De Blasi, Lijoi, & Prünster (2011). 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. in 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.

• Hjort (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.* **18**, 1259–1294.

• Favaro, Lijoi & Prünster (2011). A new estimator of the discovery probability. Tech. Report.

• 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.

• Kim & Lee (2001). On posterior consistency of survival models. *Ann. Statist.* **29**, 666–686.

• 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.

• Walker & Muliere (1997). Beta-Stacy processes and a generalization of the Pólya-urn scheme. *Ann. Statist.* **25**, 1762–1780.