On some distributional properties of Gibbs-type priors

Igor Prünster

University of Torino & Collegio Carlo Alberto

Fête Parisienne in Computation, Inference and Optimization

IHES, Paris, 20th March 2013

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

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 $n \geq 1$

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

$\Longrightarrow$ 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 $n \geq 1$

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

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

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

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

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

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

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

$\Longleftarrow$ 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 $X^{(n)} := (X_1, \ldots, X_n)$.

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

$\Longleftarrow$ 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 $\mathbf{N}_n = (N_1, \ldots, N_{K_n})$

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

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

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

$\iff$ 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

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

$\Longrightarrow$ 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

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

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

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, for which

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

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

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 (see Lijoi et al., 2007). If $\sigma = 1/2$ it reduces to the normalized inverse Gaussian process (N–IG) (Lijoi et al., 2005).

- $\sigma < 0 \implies$ 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

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

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 $\theta = r|\sigma|$.
- An interesting model (Gnedin, 2010) arises if, for $r = 1, 2, \ldots$ with $\gamma \in (0, 1)$,

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

- $\sigma < 0 \implies$ 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 $\theta = r|\sigma|$.
- An interesting model (Gnedin, 2010) arises if, for $r = 1, 2, \ldots$ with $\gamma \in (0, 1)$,

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

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!

Induced distribution on number of clusters

An alternative definition of Gibbs–type priors is as species sampling models (i.e. discrete nonparametric priors $\sum_{i\geq 1} \tilde{p}_i \delta_{Z_i}$ in which the weights p_i's and locations Z_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 $(\lozenge)$.

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 $\sum_{i\geq 1} \tilde{p}_i \delta_{Z_i}$ in which the weights p_i's and locations Z_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 $(\lozenge)$.

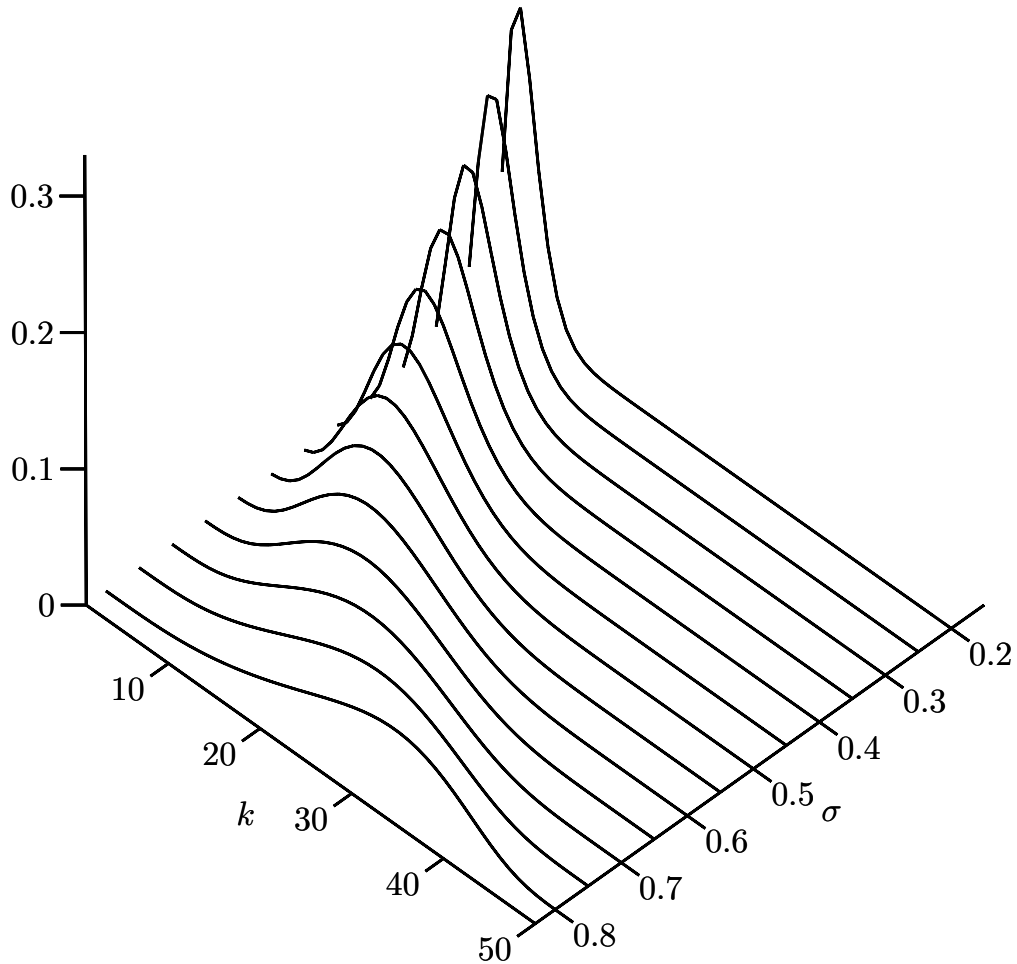
Intepretation of $(\triangle)$: probability of observing a specific sample $X_1, \ldots, X_n$ featuring j distinct observations with frequencies $n_1, \ldots, n_j$ $\implies$ exchangeable partition probability function (EPPF), a concept introduced in Pitman (1995).

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 clusters corresponding to the NGG process with $n = 50, \beta = 1$ *and* $\sigma = 0.2, 0.3, \ldots, 0.8$.

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

- 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:

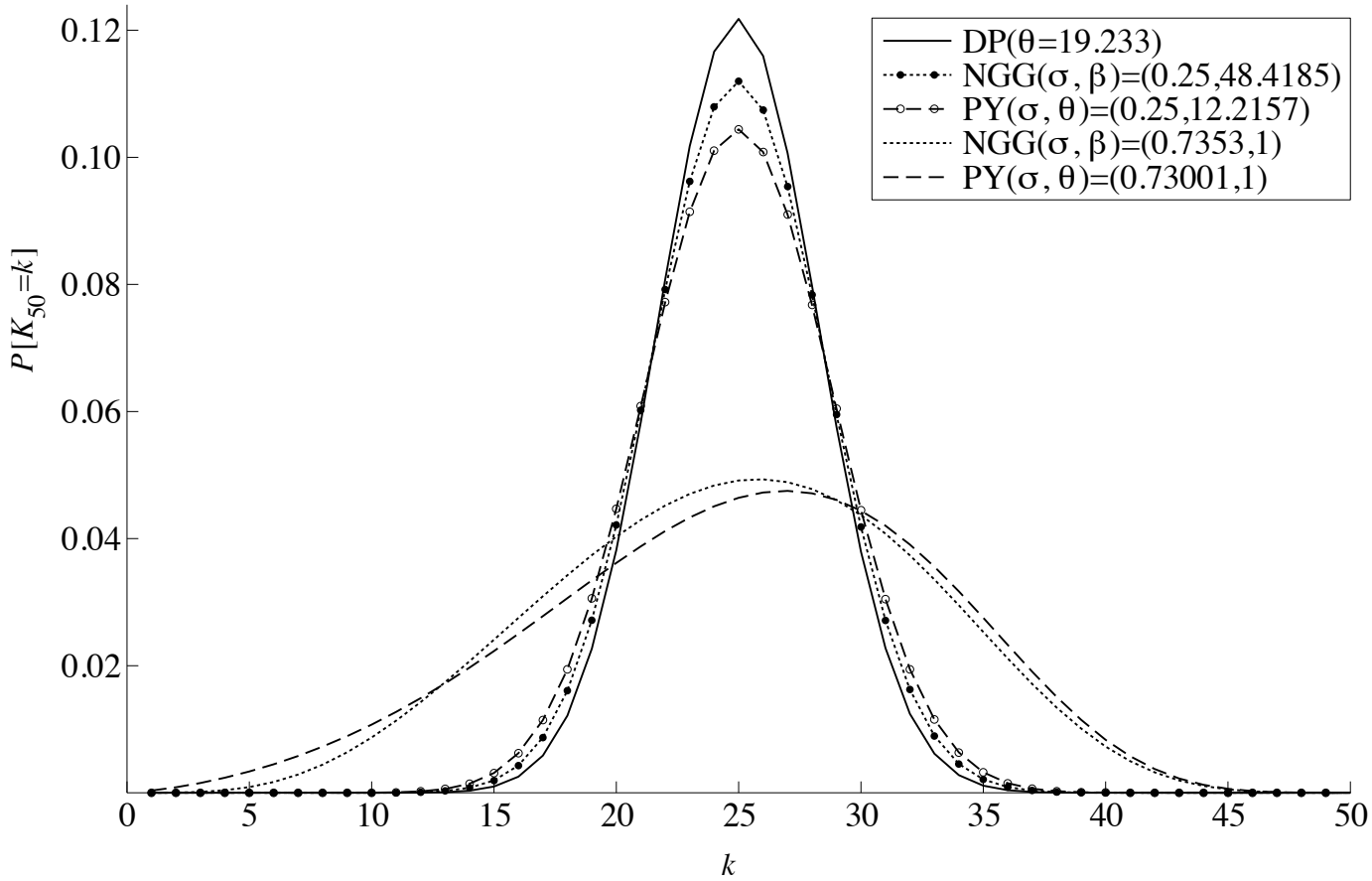
- σ controls the "flatness" (or variability) of the (prior) distribution of K_n.
- the possible second parameter (θ in the PD and β in the NGG case) controls the location of the (prior) distribution of K_n

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

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

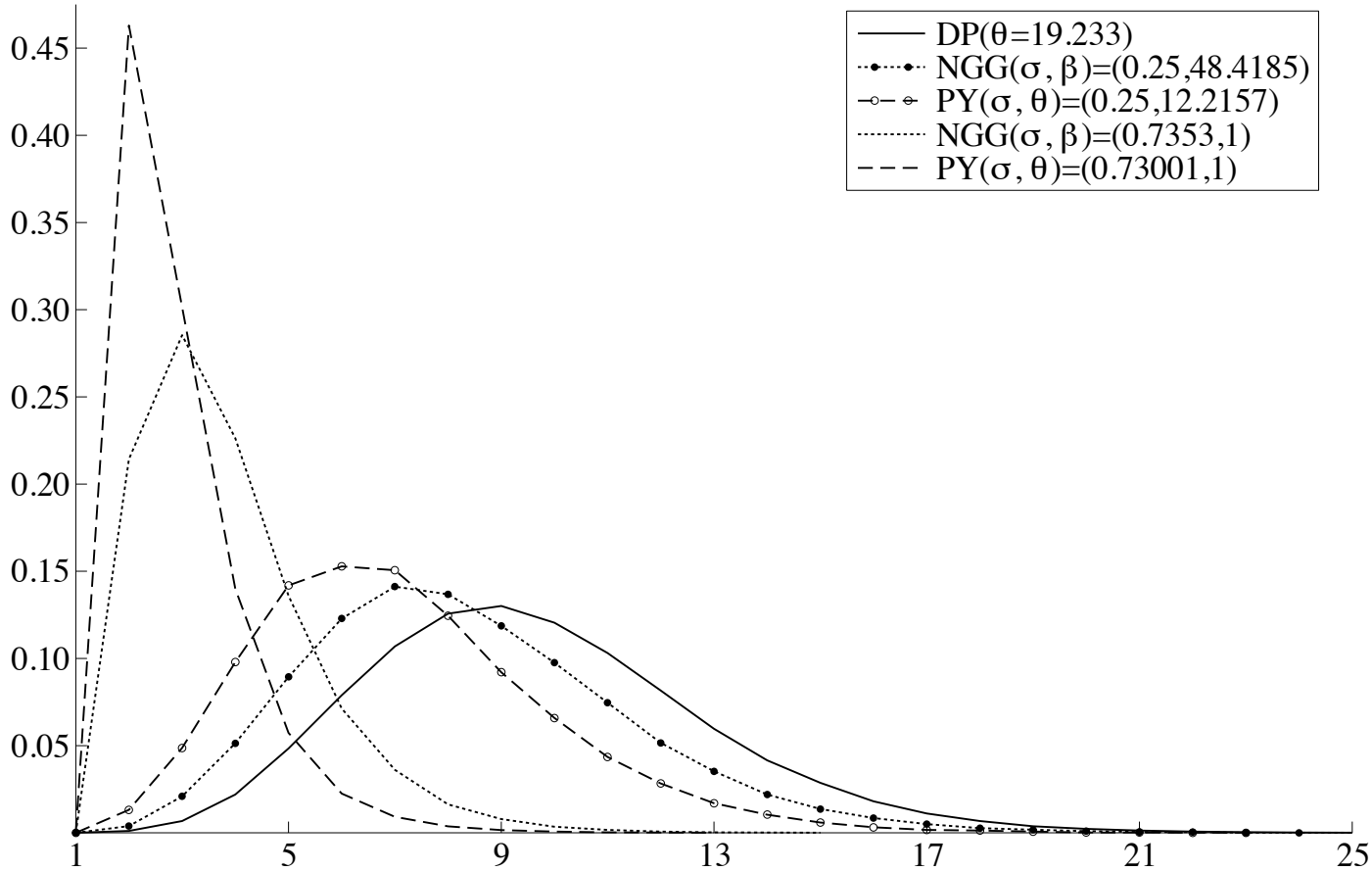
$$
\begin{aligned}
(Y_i \mid \mathrm{m}_i, \mathrm{v}_i) &\overset{\text{ind}}{\sim} \mathrm{N}(\mathrm{m}_i, \mathrm{v}_i), && i = 1, \ldots, n \\
(\mathrm{m}_i, \mathrm{v}_i \mid \tilde{P}) &\overset{\text{iid}}{\sim} \tilde{P} && i = 1, \ldots, n \\
\tilde{P} &\sim Q
\end{aligned}
$$

  with Q a Gibbs–type prior and standard specifications for P^*;
- The distribution of K_n represents the prior distribution on the number of mixture components; some summary statistics of its posterior distribution of $(K_n | Y^{(n)})$ is then used as estimate of the number of mixture components.
- 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 clusters corresponding to various choices of Gibbs–type priors with $n = 50$ *and* $\mathbb{E}(K_{50}) = 25$.

Some further properties of Gibbs–type priors

- Asymptotics for K_n:
  - In the Dirichlet case $K_n / \log n \xrightarrow{\text{a.s.}} \theta$ (Korwar and Hollander, 1973) $\Longrightarrow$ inappropriate in e.g. linguistics (Teh, 2006) and species sampling (Lijoi et al., 2007).
  - For Gibbs–type priors with $\sigma > 0$ (Gnedin and Pitman, 2006)

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

$\Longrightarrow$ by tuning σ whole spectrum of growth rates

Some further properties of Gibbs–type priors

- Asymptotics for K_n:
  - In the Dirichlet case $K_n / \log n \xrightarrow{\text{a.s.}} \theta$ (Korwar and Hollander, 1973) $\Longrightarrow$ inappropriate in e.g. linguistics (Teh, 2006) and species sampling (Lijoi et al., 2007).
  - For Gibbs–type priors with $\sigma > 0$ (Gnedin and Pitman, 2006)

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

  $\Longrightarrow$ by tuning σ whole spectrum of growth rates
- Full weak support property: Gibbs–type priors with $\sigma < 0$ & $\text{supp}(\pi) = \mathbb{N}$ or $\sigma \geq 0$ ("genuinely nonparametric Gibbs–type priors") imply that weak neighborhoods of any given distribution have *a priori* positive probability (De Blasi et al., 2013).

Some further properties of Gibbs–type priors

- Asymptotics for K_n:
  - In the Dirichlet case $K_n / \log n \xrightarrow{\text{a.s.}} \theta$ (Korwar and Hollander, 1973) $\Longrightarrow$ inappropriate in e.g. linguistics (Teh, 2006) and species sampling (Lijoi et al., 2007).
  - For Gibbs–type priors with $\sigma > 0$ (Gnedin and Pitman, 2006)

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

  $\Longrightarrow$ by tuning σ whole spectrum of growth rates
- Full weak support property: Gibbs–type priors with $\sigma < 0$ & $\text{supp}(\pi) = \mathbb{N}$ or $\sigma \geq 0$ ("genuinely nonparametric Gibbs–type priors") imply that weak neighborhoods of any given distribution have *a priori* positive probability (De Blasi et al., 2013).
- Stick–breaking representation: can be derived and the stick–breaking weights will be dependent (with the exception of the PD process and Dirichlet process for which they become independent and iid, respectively); N–IG case is the first example of explicit stick–breaking representation with dependent weights (Favaro et al., 2012).

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 $\mathbf{N} = (N_1, \ldots, N_{K_n})$ 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 $\mathbf{N}$ can also be coded by $\mathbf{M} := (M_1, \ldots, M_n)$

  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.

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

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).

- 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. In applications we will use the PD process as specific prior since it allows for completely explicit expressions.

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. In applications we will use the PD process as specific prior since it allows for completely explicit expressions.

The resulting estimators are:

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

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

and the discovery probability of a new species

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

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

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

$$\left[\overset{\text{PD case}}{=} \frac{\theta+k\sigma}{\theta+n} \frac{(\theta+n+\sigma)_m}{(\theta+n+1)_m} \right],$$

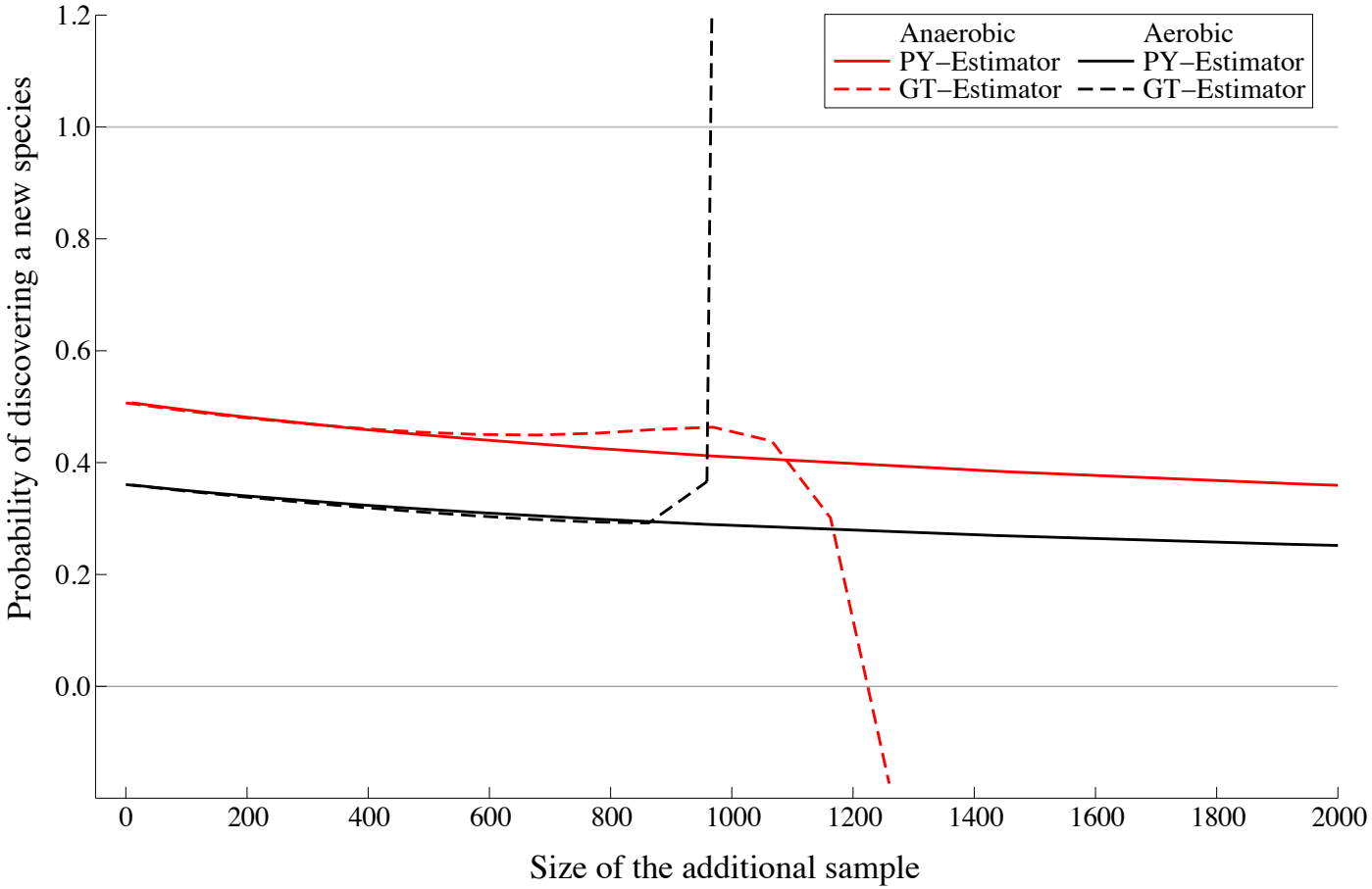
where $\mathscr{C}(m,j;\sigma,-n+k\sigma)$ is the non–central generalized factorial coefficient.

- 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.

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* $m = 1, \ldots, 2000$.

Some remarks on BNP models for 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.

Some remarks on BNP models for 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 $\sigma < 0$ (*work in progress*).

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*
>
> $$Q(A_\varepsilon|X^{(n)}) \overset{n\to\infty}{\longrightarrow} 1 \qquad a.s. - P_0^\infty$$
>
> *with A_ε a weak neighbourhood of P_0 and P_0^∞ the infinite product measure.*

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

Key quantity in this study is once again

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

which we need to converge to 0 a.s.–P_0^∞ (i.e. "washing out of the prior") to achieve consistency.

The case of discrete "true" data generating distribution P_0

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

Case I: P_0 is discrete (with either finite or infinite support points):

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.

The case of discrete "true" data generating distribution P_0

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

Case I: P_0 is discrete (with either finite or infinite support points):

> *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.*

$\Longrightarrow$ 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

Case II: P_0 is diffuse (i.e. $P_0(\{x\}) = 0$ for every $x \in \mathbb{X}$)

Erratic example: For Gnedin's model with $\sigma = -1$ and parameter $\gamma \in (0, 1)$

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

$\Longrightarrow$ concentrates around the prior guess P^* meaning that no learning at all takes place: "total" inconsistency!

The case of diffuse "true" data generating distribution P_0

Case II: P_0 is diffuse (i.e. $P_0(\{x\}) = 0$ for every $x \in \mathbb{X}$)

Erratic example: For Gnedin's model with $\sigma = -1$ and parameter $\gamma \in (0, 1)$

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

$\Longrightarrow$ concentrates around the prior guess P^* meaning that no learning at all takes place: "total" inconsistency!

One can still establish a general consistency result for diffuse P_0:

> *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 $M < \infty$*
>
> $$\frac{\pi(x+1)}{\pi(x)} \leq \frac{M}{x}.$$

Should we worry?

NO, discrete nonparametric priors are designed to model discrete distributions and should not be used to model data from diffuse distributions.

Should we worry?

NO, discrete nonparametric priors are 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!

Should we worry?

NO, discrete nonparametric priors are 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!

Some References

• De Blasi, Favaro, Lijoi, Mena, Prünster and Ruggiero (2013). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *Tech. Report.*
• De Blasi, Lijoi, & Prünster (2013). An asymptotic analysis of a class of discrete nonparametric priors. *Satist. Sinica*, in press.
• Diaconis & Freedman (1986). On the consistency of Bayes estimates. *Ann. Statist.* **14**, 1-26.
• 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* **68**,1188-96
• 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.
• 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.
• Korwar & Hollander (1973). Contribution to the theory of Dirichlet processes. *Ann. Probab.* **1**, 705-11
• Lo (1984). On a class of Bayesian nonparametric estimates. *Ann. Statist.* **12**, 351-57.
• Mao & Lindsay (2002). A Poisson model for the coverage problem with a genomic application. *Biometrika* **89**, 669-81.
• Mao (2004). Prediction of the conditional probability of discovering a new class. *J. Am. Statist. Assoc.* **99**, 1108-18.
• 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*, 985-92.