Does the full support property suffice for Bayesian nonparametrics?

Igor Prünster

Bocconi University

International FocuStat Workshop:
Bridging Parametrics and Nonparametrics

Oslo, May 23, 2017

Outline

Exchangeability

• Probabilistic statement concerning homogeneity (symmetry, analogy) among observations justifies induction

$\Longrightarrow$ Inferences are not affected by the order of the observations

• Suitable mathematical framework for inference alternative to the classical approach with "fixed and unknown probability distribution P"

$\Longrightarrow$ A sequence of observations $(X_n)_{n\geq 1}$ is exchangeable if for any $n \geq 1$ and permutation π of $(1, \ldots, n)$

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

• de Finetti's representation theorem: $(X_n)_{n\geq 1}$ is exchangeable if and only

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

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 the representation theorem in hierarchical form as

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

Remark: The representation theorem provides a neat justification for the use of prior distributions. It is the assumption of exchangeability that implies the existence of a prior distribution. In Diaconis' words "a philosophically sensational result".

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 general Bayesian framework was laid out by de Finetti in the 30's, but w.r.t. the nonparametric side, 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: introduction of Dirichlet process prior by Ferguson (1973)

Discrete nonparametric priors

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

$$\tilde{P}(\,\cdot\,) = \sum_{i=1}^{\infty} \tilde{p}_i \delta_{Z_i}(\,\cdot\,),$$

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

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

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

Throughout the weights $\tilde{p}_i$'s are assumed independent from the locations Z_i's, which are i.i.d. from P^*:

- $\mathbb{E}[\tilde{P}(\,\cdot\,)] = \mathbb{E}[\sum_{i=1}^{\infty} \tilde{p}_i \delta_{Z_i}(\,\cdot\,)] = P^*(\,\cdot\,) =$ "prior guess at the shape of the data generating distribution"
  e.g. one can set P^* to be a $N(0,1)$.

- In Hjort's words Q represents a "nonparametric envelope around P^*"

Support of a discrete nonparametric prior

Recall that:

- If ρ is a probability measure on a Polish (i.e. complete and separable metric) space equipped with the Borel σ-field, then supp(ρ) coincides with the smallest closed set of probability 1.
- If $\mathbb{X}$ is Polish, then $\mathscr{P}$ endowed with the Borel σ-algebra for the topology of weak convergence is also Polish.
- If ρ is a probability measure on $\mathscr{P}$, supp(ρ) is often referred to as weak support.

Under typically mild technical conditions, the support of a discrete nonparametric prior Q on $\mathscr{P}$ as defined previously coincides with

$$supp(Q) = \{P \in \mathscr{P} : supp(P) \subseteq supp(P^*)\}$$

In particular, if $supp(P^*) = \mathbb{X}$, then $supp(Q) = \mathscr{P}$.
$\Longrightarrow$ "full weak support"

Uses of discrete nonparametric priors

Discrete nonparametric priors are popular tools for addressing the following large classes of problems:

1. Species sampling: $\tilde{P} = \sum_{i=1}^{\infty} \tilde{p}_i \delta_{Z_i}$ used as a model for the species' distribution within a population, where the $\tilde{p}_i$'s interpreted as species proportions. Hence, a sample $X^{(n)} = (X_1, \ldots, X_n)$ from $\tilde{P}$ is read as:
   - 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: $\tilde{P}$ placed at a latent level of a hierarchical mixture (with f a kernel, typically Gaussian)

$$\begin{aligned} Y_i \,|\, \theta_i &\overset{\text{ind}}{\sim} f(\,\cdot\,|\theta_i) && i = 1, \ldots, n \\ \theta_i \,|\, \tilde{P} &\overset{\text{iid}}{\sim} \tilde{P} && i = 1, \ldots, n \\ \tilde{P} &\sim Q. \end{aligned}$$

   $\Longrightarrow$ the discreteness of $\tilde{P}$ allows for clustering: K_n, i.e. the number of distinct θ_i's, represents the prior number of mixture components

   $\Longrightarrow$ many successful applications can be traced back to this idea due to Lo (1984) where the mixture of Dirichlet process is introduced.

Dirichlet process via predictive distributions

Problem: Assume $(X_n)_{n\geq 1}$ is an exchangeable sequence.

- 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 $N_1, \ldots, N_{K_n}$;
- Prior guess at law of any of the X_i's is P^*;
- The strength of the prior belief is measured by a parameter $\theta > 0$.

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]

Probability of discovering a new species

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

Fundamental Characterization:
According to $(*)$ discrete $\tilde{P}$ classified in 3 categories :

(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;

(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

$\tilde{P}$ 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

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

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

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

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

Remark. Gnedin and Pitman (2006) introduced Gibbs–type exchangeable random partitions and the extension to Gibbs–type random probability measure is immediate.

Predictive characterization of the Pitman–Yor process

With $\sigma \geq 0$ 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)

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

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

Who are the members of this class of priors?

Characterization of Gibbs–type priors according to the value of σ (Gnedin and Pitman, 2006):

- $\sigma = 0 \iff$ Dirichlet process or Dirichlet process mixed over its total mass parameter $\theta > 0$;
- $0 < \sigma < 1 \iff$ 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.
- $\sigma < 0 \implies$ mixtures of symmetric k–variate Dirichlet distributions

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

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 $\theta = r|\sigma|$.
  - 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.

Full weak support property

It is well know that the Dirichlet process has full weak support.

Also Gibbs–type priors with

- $\sigma \geq 0$
- $\sigma < 0$ & $\text{supp}(\pi) = \mathbb{N}$

have full weak support $\implies$ "genuinely nonparametric Gibbs–type priors"

Same question from different angles:

- Is the full weak support property sufficient for guaranteeing the required modeling and inferential flexibility?
- Does the distribution of functionals of interest depend on features of the discrete nonparametric prior different from the base measure P^*?
- Is it equivalent to use the Dirichlet process or any other Gibbs–type prior as long as the base measures P^* are the same?
- Is it worth framing a problem within the generality of Gibbs–type priors and then, according to the problem select the one that fits best?

A first functional of interest: K_n

Gibbs–type priors induce a random partition of the form

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

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

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

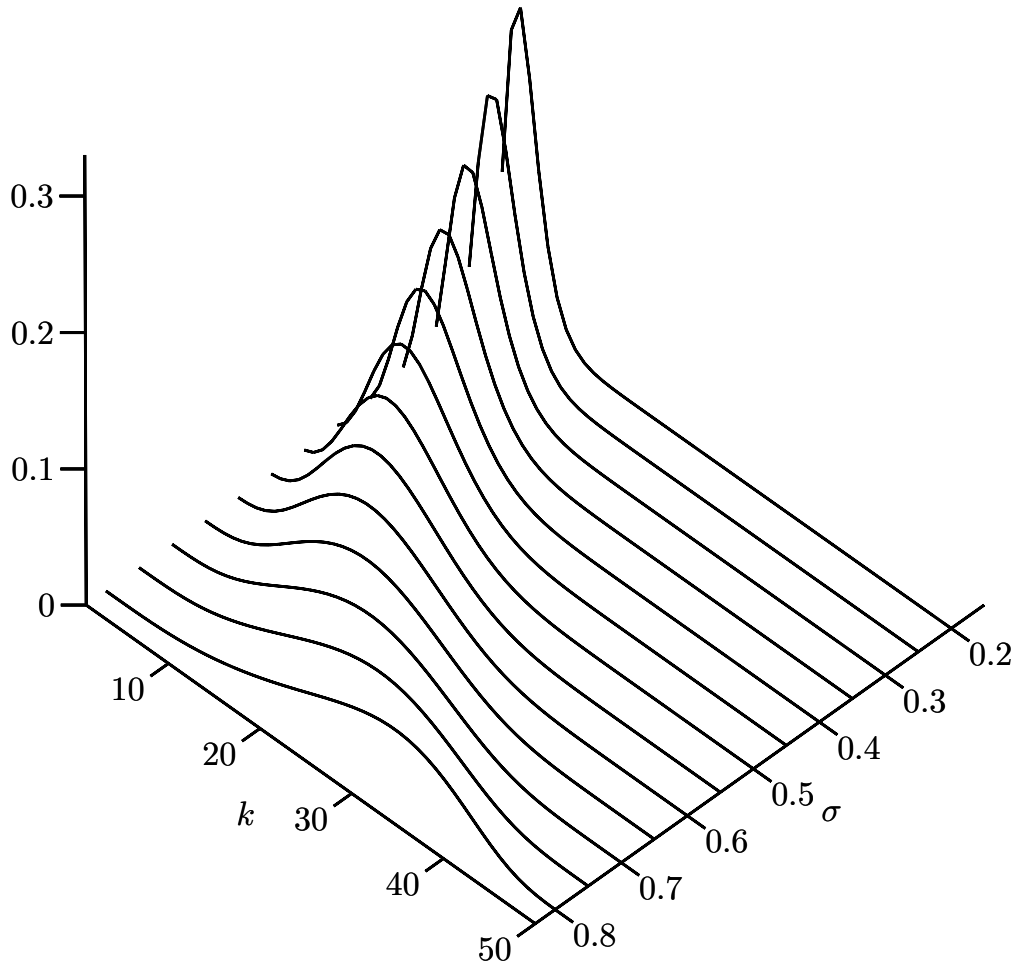
Consequently, one obtains the (prior) distribution of the number of clusters by summing over all possible partitions of a given size

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

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

$\Longrightarrow$ From an inferential perspective interest is in the posterior distribution of K_n given the data

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 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

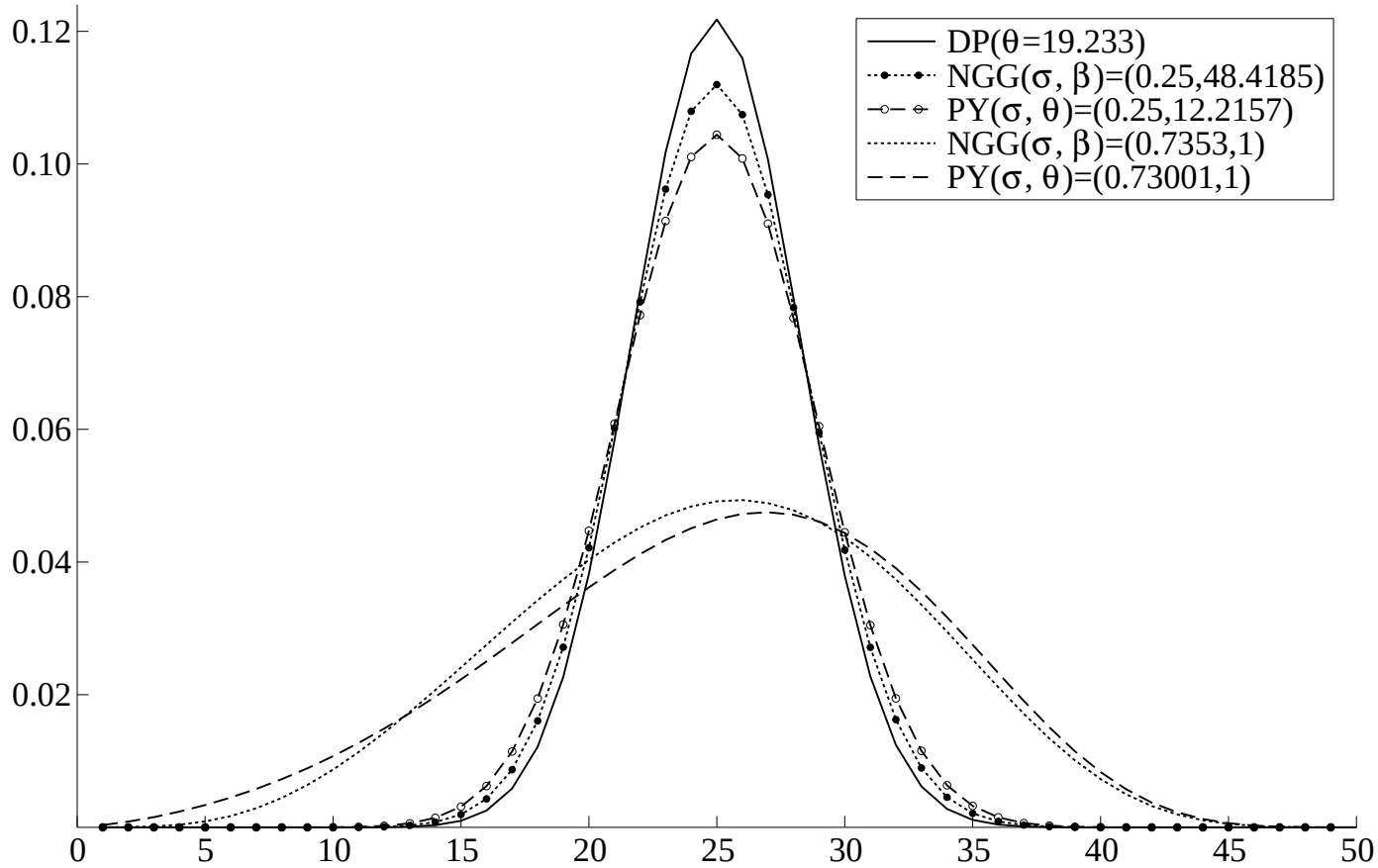
Comparative example of different Gibbs–type priors:

- $n = 50$ 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 $\theta = 19.233$;
  - PY processes with $(\sigma, \theta) = (0.73001, 1)$ and $(\sigma, \theta) = (0.25, 12.2157)$;
  - NGG processes with $(\sigma, \beta) = (0.7353, 1)$ and $(0.25, 48.4185)$.

$\implies$ 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

- $n = 50$ 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

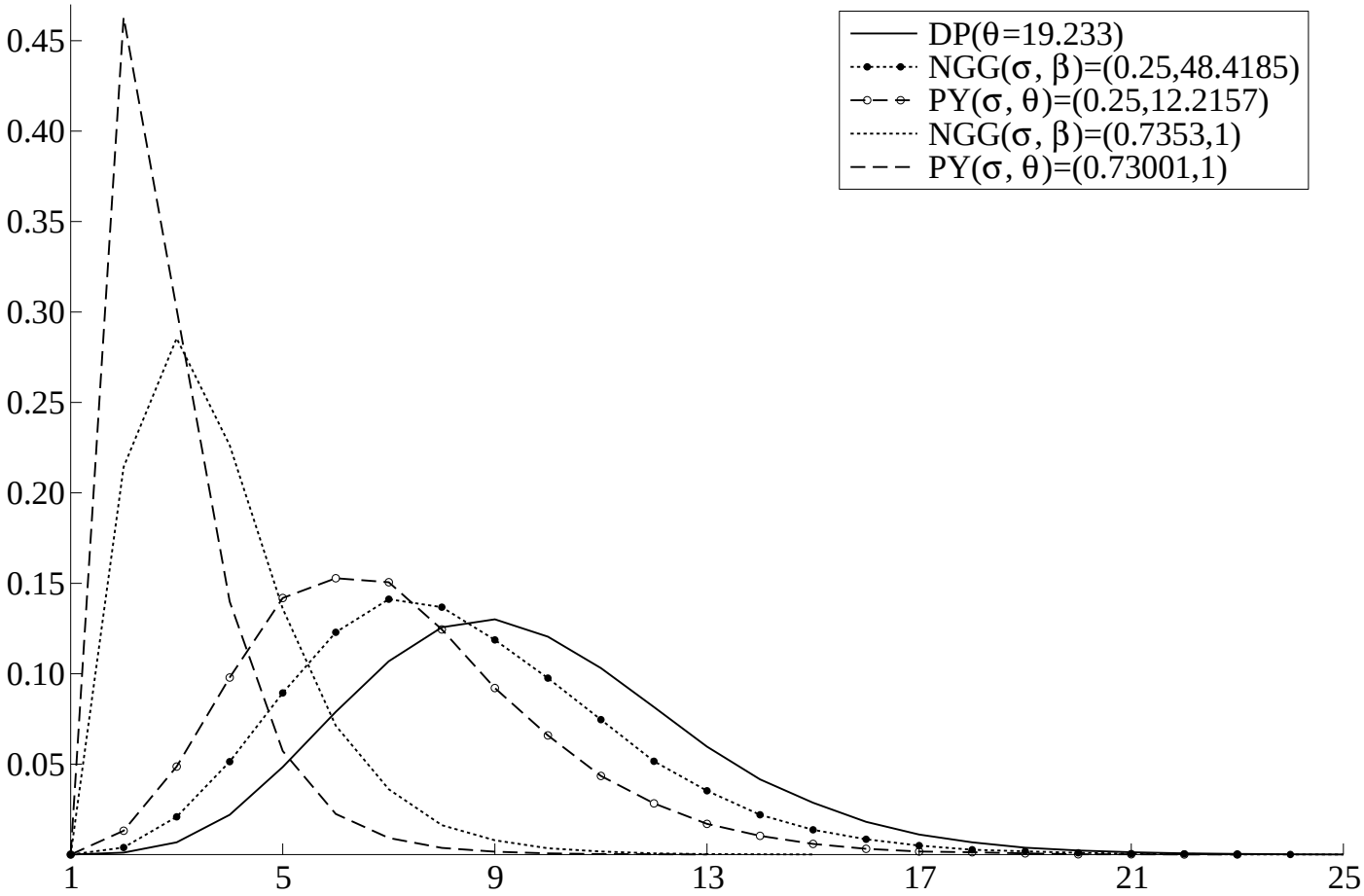
$$
\begin{aligned}
(Y_i \mid \mathrm{m}_i, \mathrm{v}_i) &\overset{\text{ind}}{\sim} \mathrm{N}(\mathrm{m}_i, \mathrm{v}_i), && i = 1, \dots, n \\
(\mathrm{m}_i, \mathrm{v}_i \mid \tilde{P}) &\overset{\text{iid}}{\sim} \tilde{P} && i = 1, \dots, n \\
\tilde{P} &\sim Q
\end{aligned}
$$

  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.
- Goal: posterior distribution of K_n given the data $Y^{(n)}$.
- 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?

$\Longrightarrow$ a large σ allows to overcome misspecification.

Posterior distribution of the number of clusters



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

Data structure in species sampling problems

- $X^{(n)}$ = basic sample of draws from a population containing different species (plants, genes, animals,...). Information:
  - ◇ sample size n and number of distinct species in the sample K_n;
  - ◇ a collection of frequencies $\boldsymbol{N} = (N_1, \ldots, N_{K_n})$ 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 $X^{(n)} = (X_1, \ldots, X_n)$, the inferential goal consists in prediction about various features of an additional sample $(X_{n+1}, \ldots, X_{n+m})$.

Carnap (1950) provides a taxonomy of the varieties of inductive inference:

> *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

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

- m–step prediction

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

A second functional of interest: the discovery probability

Discovery probability $\Longrightarrow$ conditional on the basic sample $X^{(n)}$, 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 $X_{n+1}, \ldots, X_{n+m}$.

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

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

Frequentist nonparametric estimators

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

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

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

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

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

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

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

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 PY 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

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

and the discovery probability of a new species

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

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

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

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

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

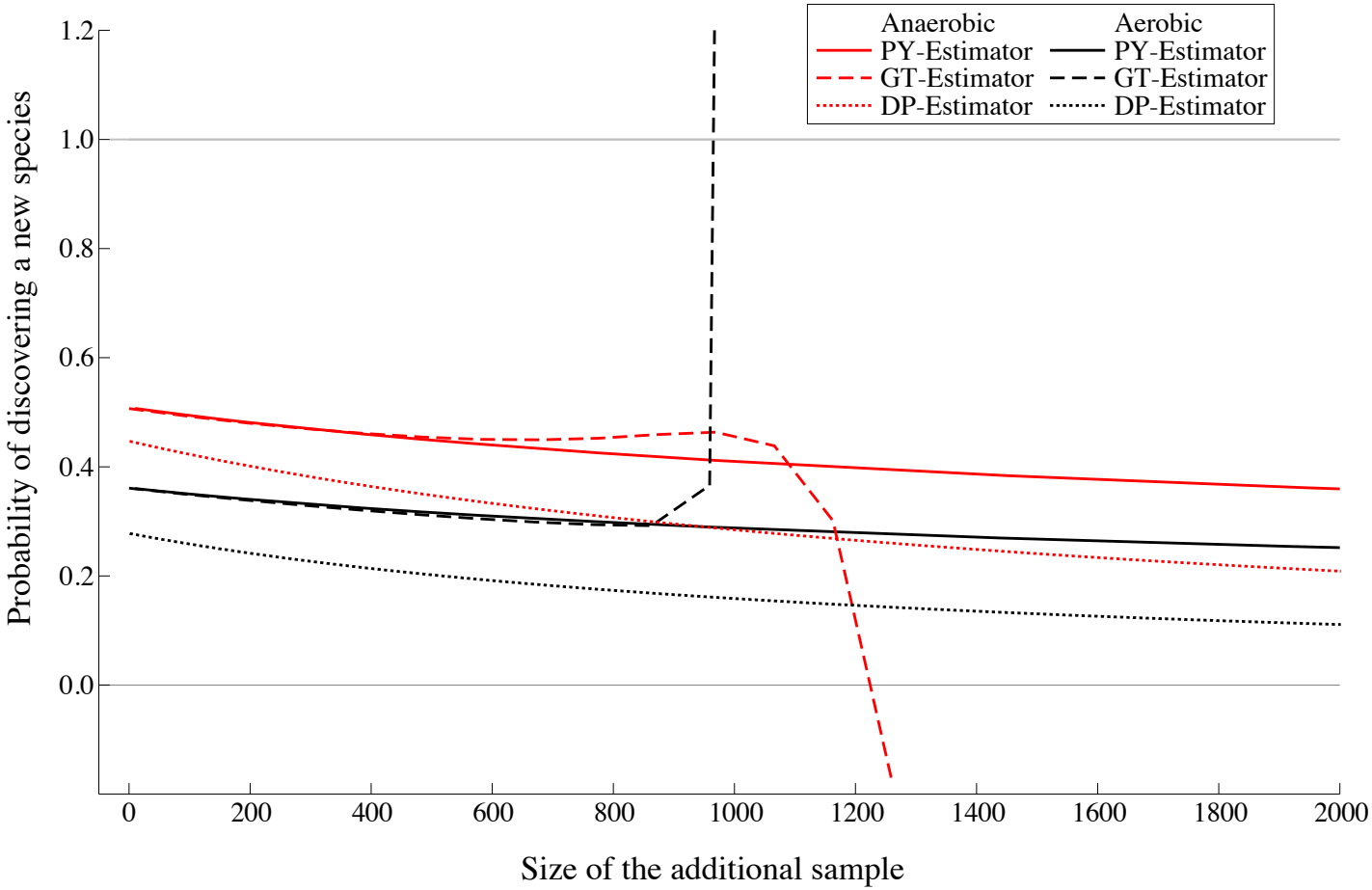
- BNP estimator for the probability of discovering a species with frequency r at the (n+m+1)–th sampling step

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

$$\left[\overset{\text{PY\_case}}{=} \sum_{i=1}^{r} m_i (i-\sigma)_{r+1-i} \binom{m}{r-i} \frac{(\theta+n-i+\sigma)_{m-r+i}}{(\theta+n)_{m+1}} \right.$$

$$\left. + (1-\sigma)_r \binom{m}{r} \frac{(\theta+k\sigma)(\theta+n+\sigma)_{m-r}}{(\theta+n)_{m+1}} \right]$$

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), DP process and PY process estimators of the probability of discovering a new gene at the* $(n+m+1)$*–th sampling step for* $m = 1, \ldots, 2000$.

Asymptotic behaviour of the number of clusters K_n

- A priori asymptotic behaviour of 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), certain graph structures (Caron, 2012; Caron & Fox, 2017) and species sampling.
  - 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

- A posteriori asymptotic behaviour: Consider $K_m^{(n)} := K_{m+n} - K_n$ ='number of new species to be recorded in the additional sample of size m' for n fixed and $m \to \infty$
  - Dirichlet case conditional on $X^{(n)}$: $K_m^{(n)}/\log m \xrightarrow{\text{a.s.}} \theta$ as $m \to \infty$.
  - PY case conditional on $X^{(n)}$:

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

  with $Z_{n,j}$ the product of a beta and transformed positive stable r.v. $\Longrightarrow$ asymptotic credible intervals via quantiles of limiting r.v.

General remarks on BNP models

- Full weak support property is almost a "necessary condition" for a discrete nonparametric prior but far from being "sufficient"
- The full weak support property of the Dirichlet process, combined with its weak consistency (even for a continuous "true" data generating distribution!), led many to think that the Dirichlet process was an all-purpose machine!
- However, BNP models correspond to large probabilistic models in which all functionals of potential interest are implicitly (and often unconsciously!) modeled jointly when assigning the prior over $\mathscr{P}$.
- The joint modeling ensures coherence among the distributions of all objects of interest. This, however, does not mean that their distributions have the desired properties and this cannot always be fixed by tuning the base measure or adding hyperpriors.
  $\Longrightarrow$ The distribution of the $\tilde{p}_i$'s does matter!

Addendum: a different look at the nonparametric envelope

Let $\tilde{P}$ be a PY process on some functional space $\mathscr{F}$ with a spike & slab base measure

$$P^a(\cdot) = z\,P^*(\cdot) + (1-z)\,\delta_{f_0}(\cdot),$$

P^* a diffuse probability on $\mathscr{F}$ and $f_0 \in \mathscr{F}$ is some parametric function, which represents "a strong prior guess for $(1-z)$% of the data".

The realizations of $X^{(n)}$ will now be functions f_i^* and the predictive distributions are of the form

$$P[X_{n+1} \in \cdot \mid X^{(n)}] = p_{\text{new}} P^*(\cdot) + p_0 \delta_{f_0}(\cdot) + \sum_{i: X_i^* \neq x_0} p_i \delta_{f_i^*}(\cdot) \qquad (*)$$

$$p_{\text{new}} = \frac{z\,\sigma}{\theta+n}\frac{\zeta_{n_j,k+1}}{\zeta_{n_j,k}}, \quad p_0 = \frac{1}{\theta+n}\frac{\zeta_{n_j+1,k}}{\zeta_{n_j,k}}, \quad p_i = \frac{n_i-\sigma}{\theta+n}, \qquad i \neq j$$

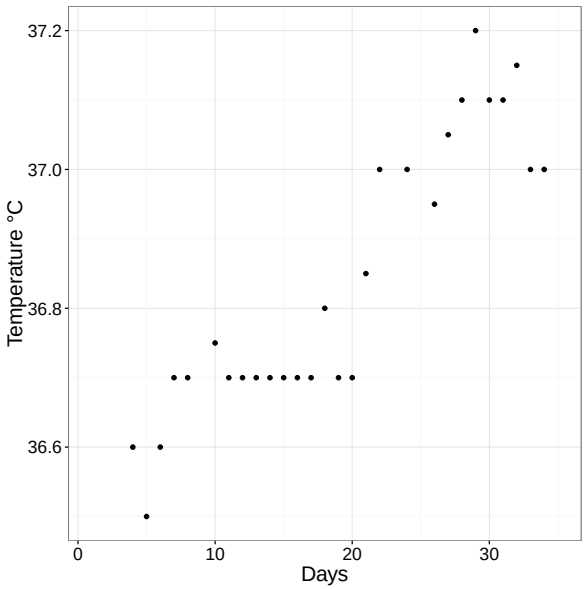
$$\zeta_{n_j,k} = \sum_{i=1}^{n_j} (1-z)^i \mathscr{C}(n_j, i; \sigma)\Gamma(\theta/\sigma + k + i),$$

Clearly, unless $f_1^*, f_2^*, \ldots, f_k^* \neq f_0$,

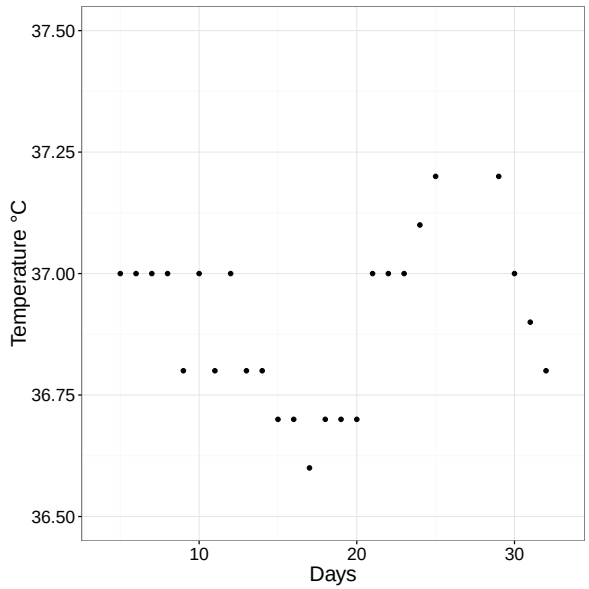
$$P[X_{n+1} \in \cdot \mid X^{(n)}] \neq \frac{\theta + k\sigma}{\theta+n} P^a(\cdot) + \sum_{i=1}^{k} \frac{n_i - \sigma}{\theta+n} \delta_{f_i^*}(\cdot)$$

Illustration on BBT curves

- Daily measurements of Basal Body Temperature (BBT)
- 1118 non-conception cycles from $n = 157$ women of reproductive age
- BBT curves of healthy women follow a biphasic trajectory
- Unhealthy women might have highly irregular BBT curves
- Goals: inference on curves' shape and clustering.



Observation 1: S-shaped curve

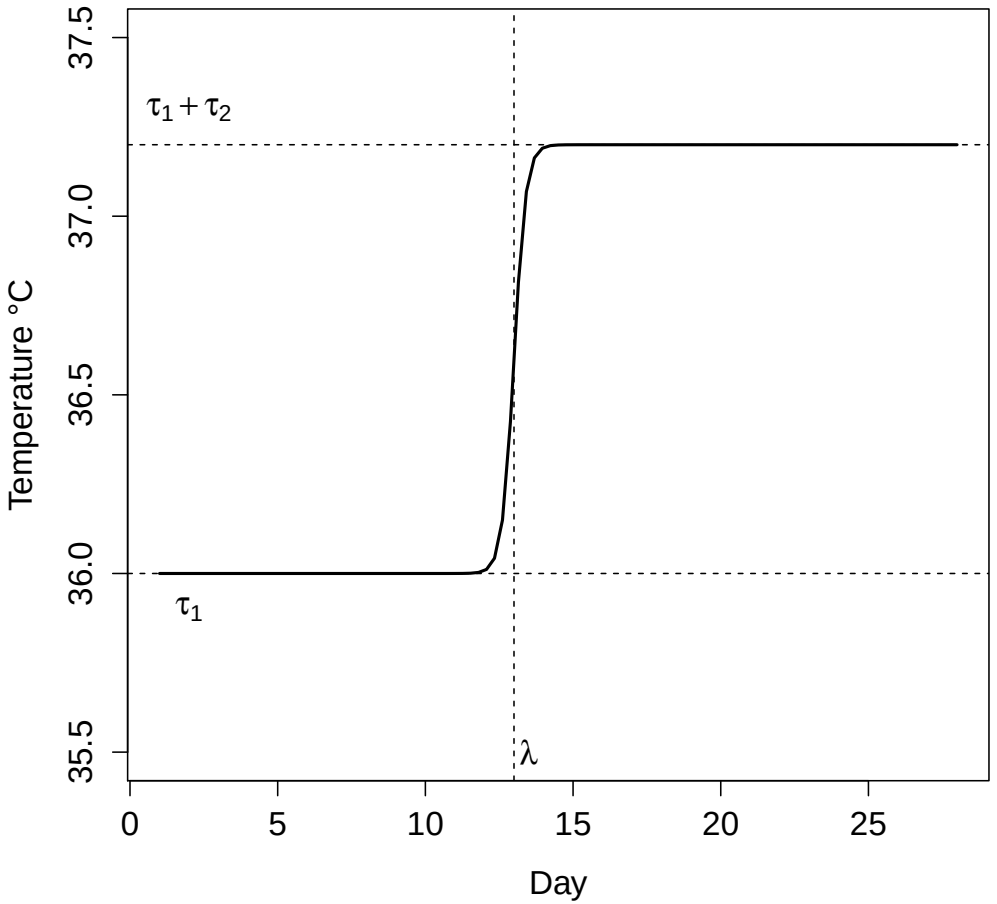


Observation 2: Irregular curve

BBT curve of healthy women

To describe the biphasic trajectory we use

$$f_0(t;\tau_1,\tau_2,\omega,\lambda) = \tau_1 + \tau_2 \left(\frac{\exp\{\frac{t-\lambda}{\omega}\}}{1+\exp\{\frac{t-\lambda}{\omega}\}} \right)$$



Parameters:

τ_1: hypothermia during the follicular phase

τ_2: increase of temperature following ovulation

ω: sharpness of temperature increase

λ: moment of ovulation

Model specification

$$y_{ij}(t) = \tau_{1ij} + \tau_{2ij} f_{ij}\left(\frac{t-\lambda_{ij}}{\omega_{ij}}\right) + \epsilon_{ij}(t)$$

$$\epsilon_{ij}(t) \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$$

$$f_{ij} \mid \tilde{P} \overset{\text{iid}}{\sim} \tilde{P}$$

$$\tilde{P} \sim \mathrm{PY}(\theta, \sigma; P^a)$$

where:

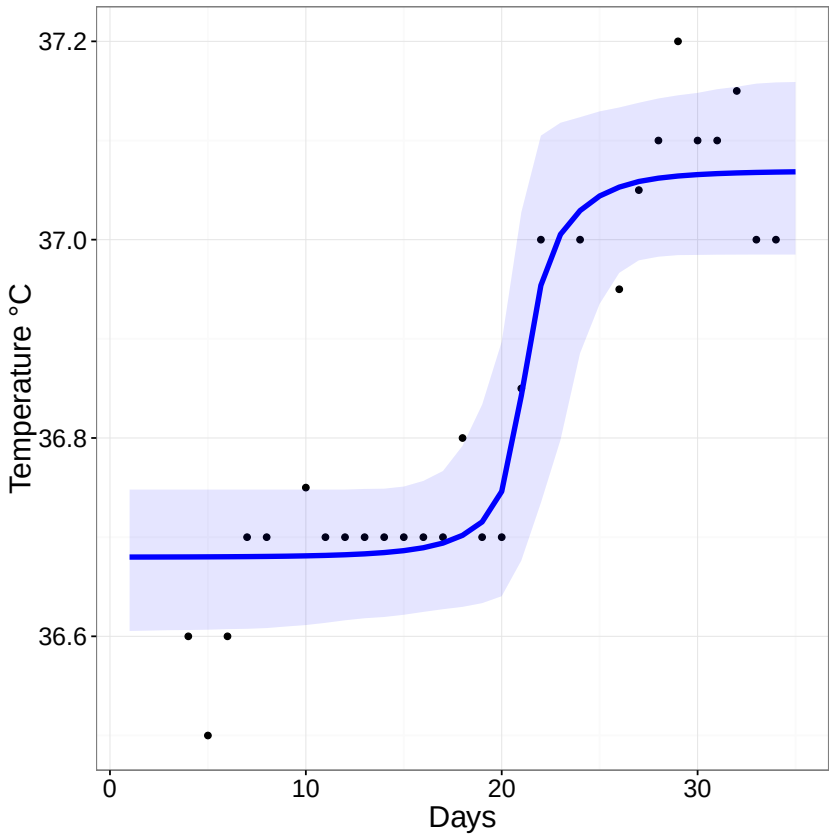
- the parameters τ_{1ij}, τ_{2ij}, λ_{ij} and ω_{ij} refer to cycle j of woman i
- ϵ_{ij} are independent measurement errors
- f_{ij} is a random function
- $\tilde{P}$ is distributed as a PY with spike &slab measure P^a

Model hyperparameters:

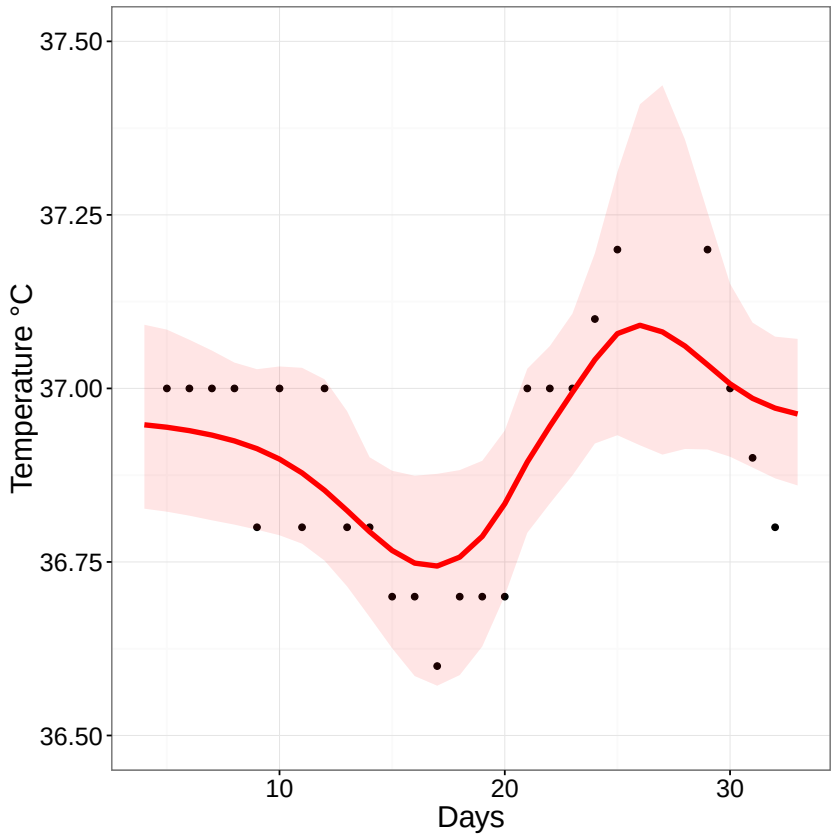
$$\sigma^{-2} \sim \mathrm{Ga}(a_\sigma, b_\sigma) \qquad (\tau_{1ij}, \tau_{2ij}) \sim \mathrm{N}(\alpha_i, \Omega)$$

$$\alpha_i \overset{\text{iid}}{\sim} \mathrm{N}(\alpha, R) \qquad \lambda_{ij} \sim \mathrm{U}(b_{ij}+10, b_{ij}+20) \qquad \omega_{ij} \overset{\text{iid}}{\sim} \mathrm{Ga}(1/2, 1)$$

Estimated curves

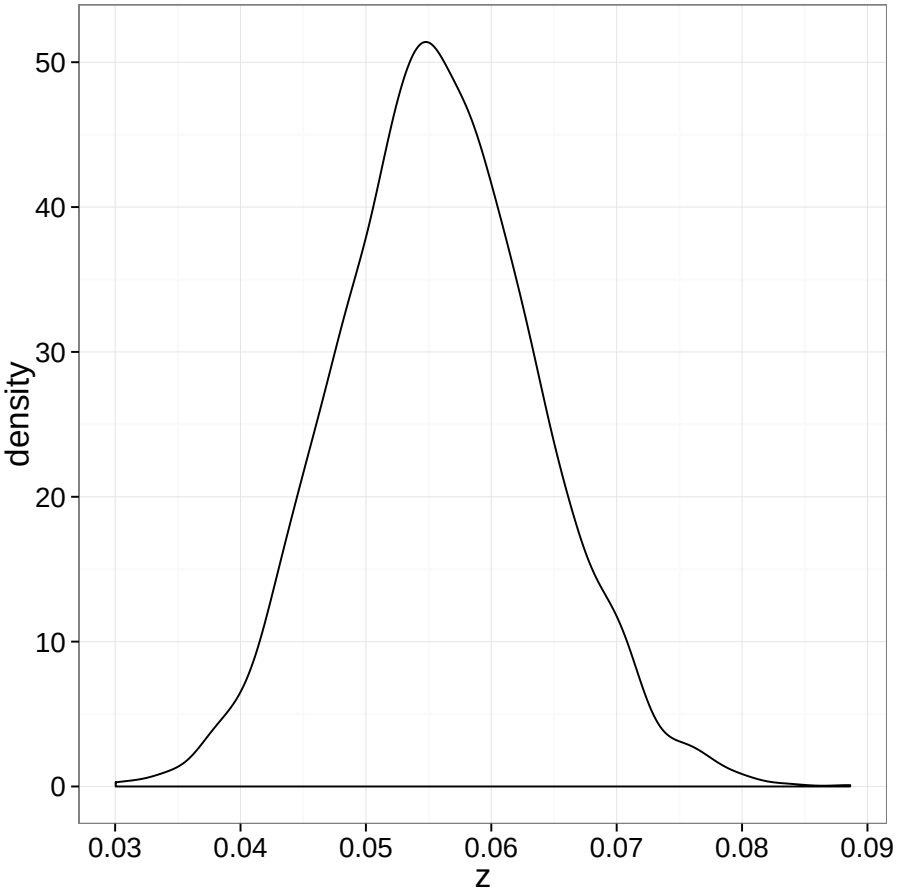


Observation 1: S-shaped curve



Observation 2: Irregular curve

Inference on z



Posterior distribution of z

- a priori $z \sim \mathrm{U}(0,1)$
- $1 - \hat{z} = 0.944$
- 94.5% are considered S-shaped
- the "most likely" partition of $n = 1118$ (according to a criterion proposed in Dahl, 2006)

cluster	S-shaped	II	III	IV	V	VI	VII	VIII
size	1064	20	14	10	5	2	2	1

Some References

• Canale, Lijoi, Nipoti & Prünster (2016). On the Pitman–Yor process with spike and slab prior specification. *Tech. Report*.
• Caron (2012). Bayesian nonparametric models for bipartite graphs. NIPS'2012.
• Caron & Fox (2017). Sparse graphs using exchangeable random measures. JRSS B, forthcoming.
• Carnap (1950). *Logical Foundations of Probability*. U. Chicago Press, Chicago.
• Dahl (2006). Model-Based Clustering for Expression Data via a Dirichlet Process Mixture Model. In *Bayesian Inference for Gene Expression and Proteomics*, CUP.
• 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.
• Favaro, Lijoi, Mena & Prünster (2009). Bayesian nonparametric inference for species variety with a two parameter PD process prior. JRSS B 71, 993-1008.
• 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.
• 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.
• Lijoi, Mena & Prünster (2007). Bayesian nonparametric estimation of the probability of discovering a new species. *Biometrika* 94, 769-786.
• Lijoi, Mena & Prünster (2007). Controlling the reinforcement in Bayesian nonparametric mixture models. *JRSS B* 69, 715-740.
• Lo (1984). On a class of Bayesian nonparametric estimates. *Ann. Statist.* 12, 351-57.
• Lo (1991). A characterization of the Dirichlet process. *S&P Lett.* 12, 185-7.
• 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.
• Pitman (1995).Exchangeable and partially exchangeable random partitions. *Prob. Th. and Rel. Fields* 102, 145–158.
• Pitman and Yor (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* 25, 855-900.
• Regazzini (1978). Intorno ad alcune questioni relative alla definizione del premio secondo la teoria della credibilità. *Giorn. Istit. Ital. Attuari* 41, 77–89.
• Teh (2006). A Hierarchical Bayesian Language Model based on Pitman-Yor Processes. *Coling/ACL 2006*, 985-92.