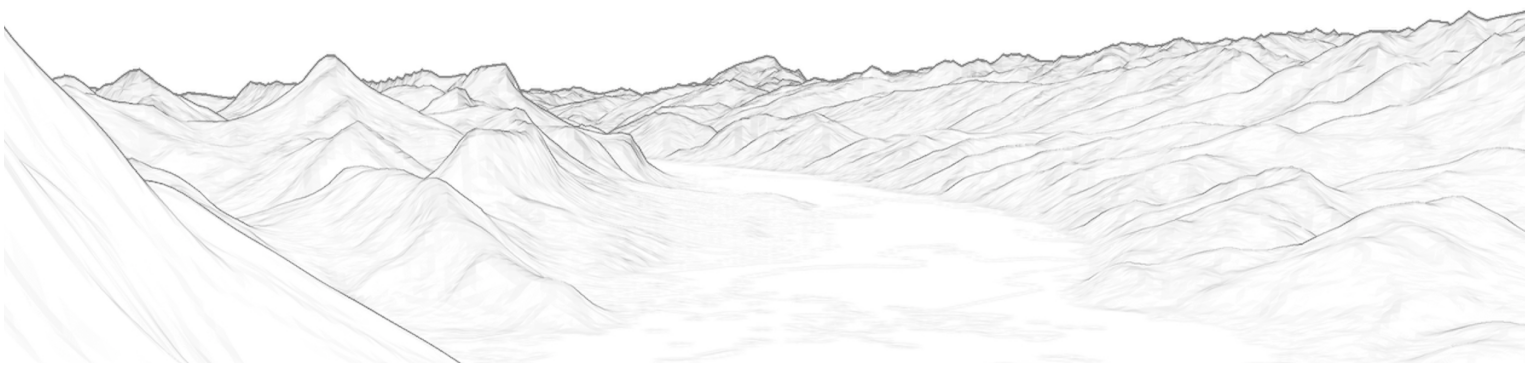
Introduction to Bayesian nonparametrics

✉ julyan.arbel@inria.fr ☜ www.julyanarbel.com

Inria Grenoble, France

Rencontres Statistiques de Rochebrune
26-28 March, 2018, Megève, France



Acknowledgement and useful links I

My slides are inspired by the following introductions to Bayesian nonparametric approaches that I found myself very useful:

- Botond Szabo's tutorial introduction
- Kurt Miller's tutorial introduction
- Peter Orbanz' tutorials webpage, as well as his lecture notes
- Yee Whye Teh's tutorial at MLSS 2011
- Mike Jordan's tutorial at NIPS 2005

Acknowledgement and useful links II

I also have some handwritten lecture notes for a Master/PhD course on Bayesian nonparametrics

Thanks Michał Lewandowski for typing help.

Table of Contents

Table of Contents

Parametric versus nonparametric

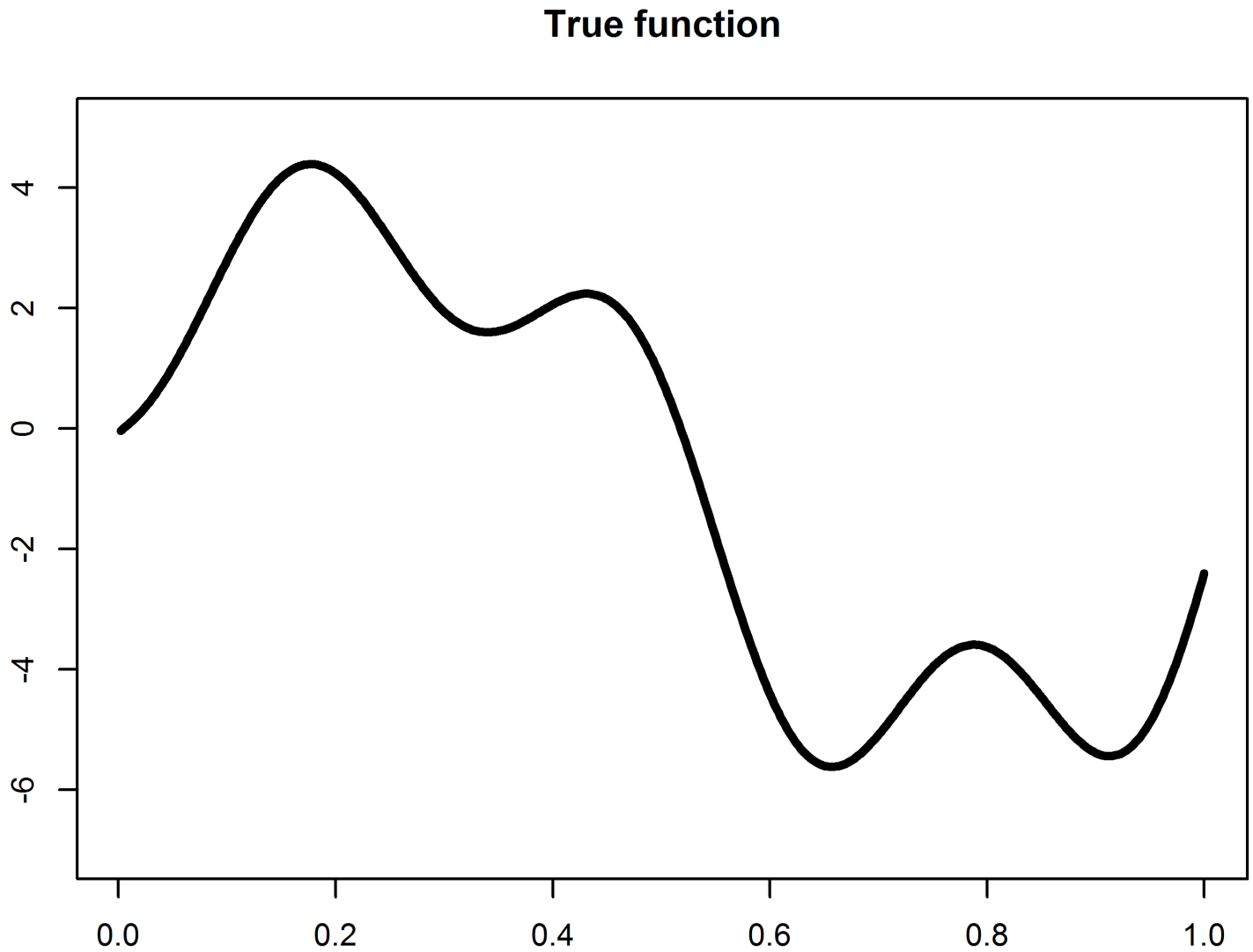
Parametric models

- Finite and fixed number of parameters
- Number of parameters is independent of the dataset

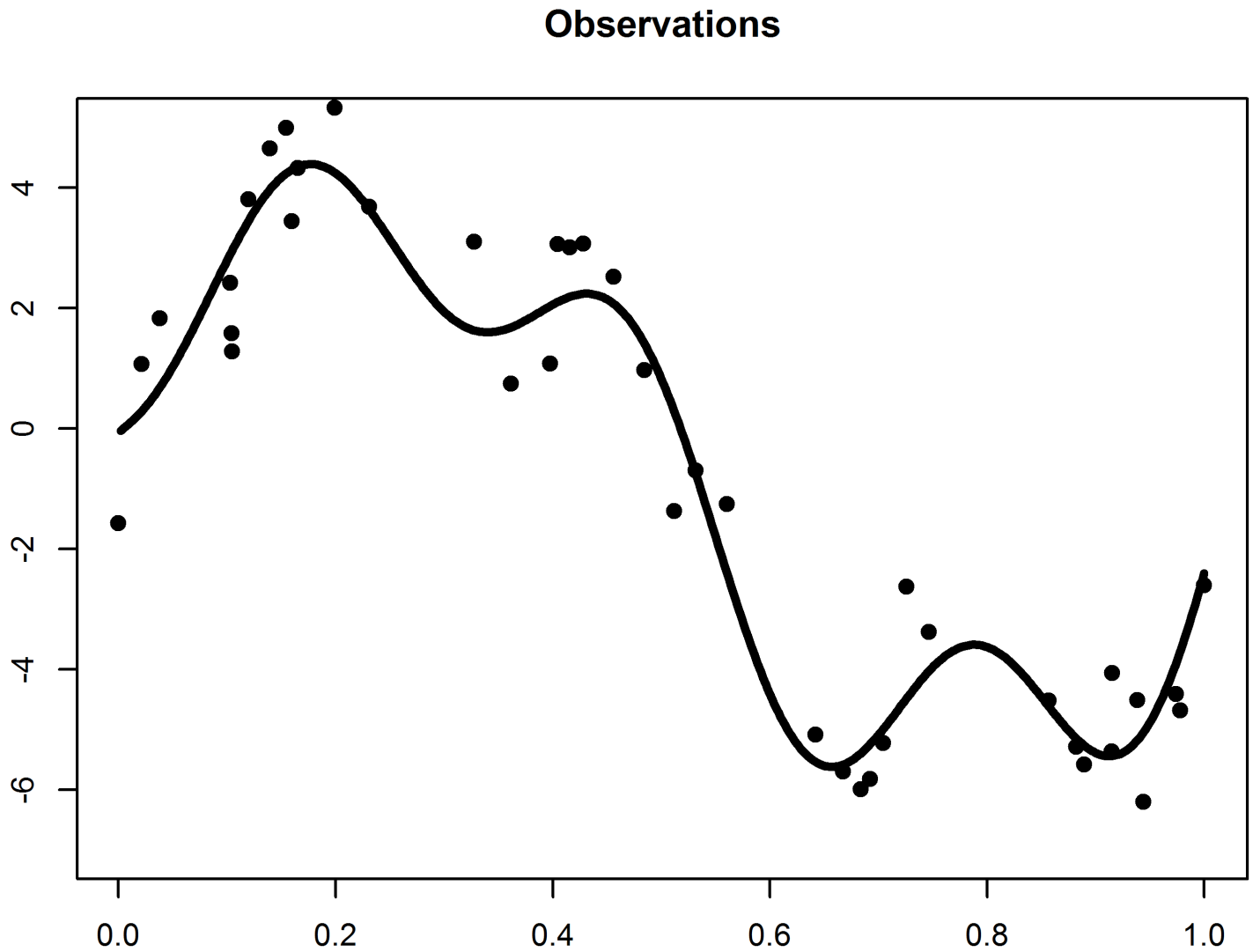
Nonparametric models

- Do have parameters
- Can be understood as having an infinite number of parameters
- Can be understood as having a random number of parameters
- Number of parameters can grow with the dataset

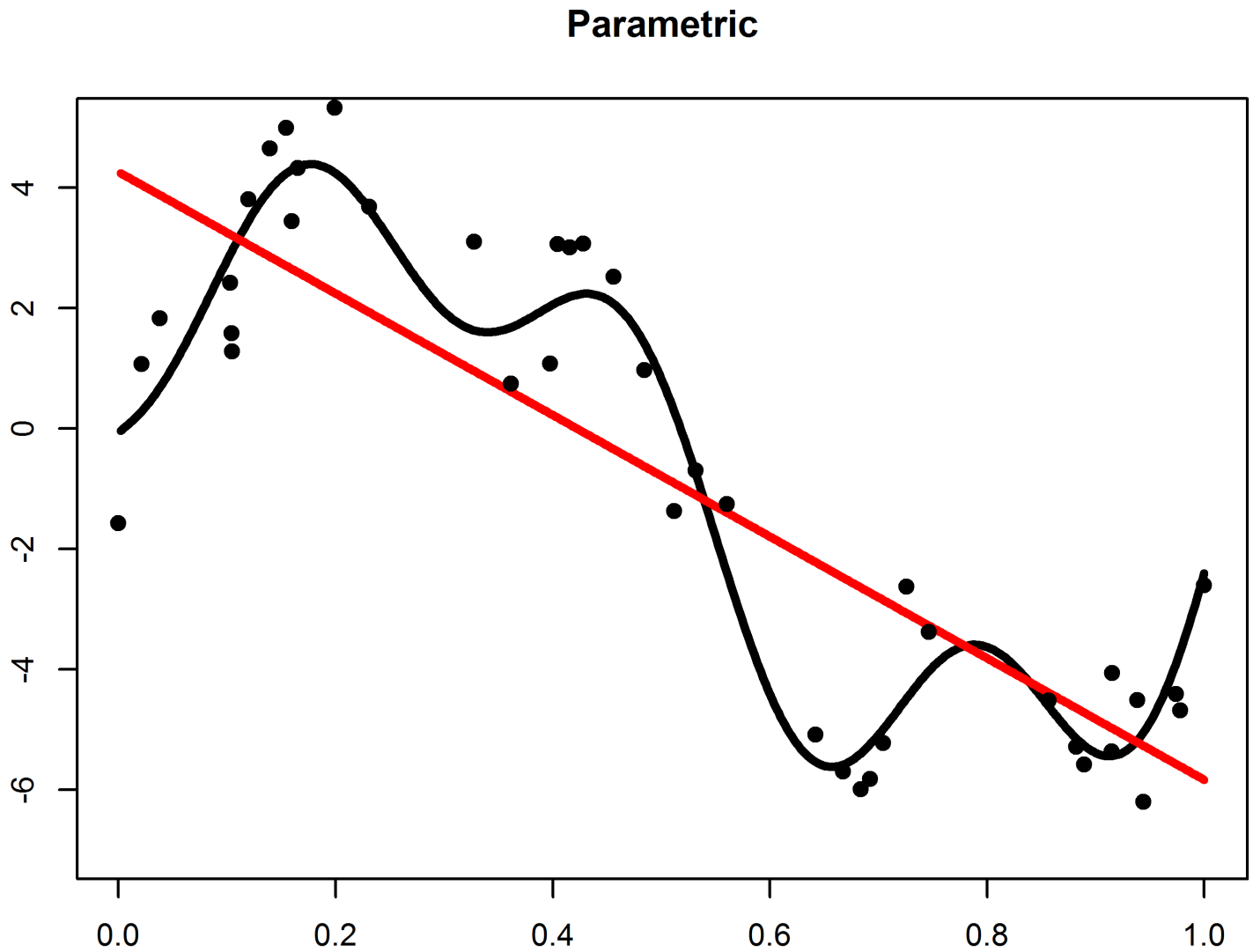
Underlying function



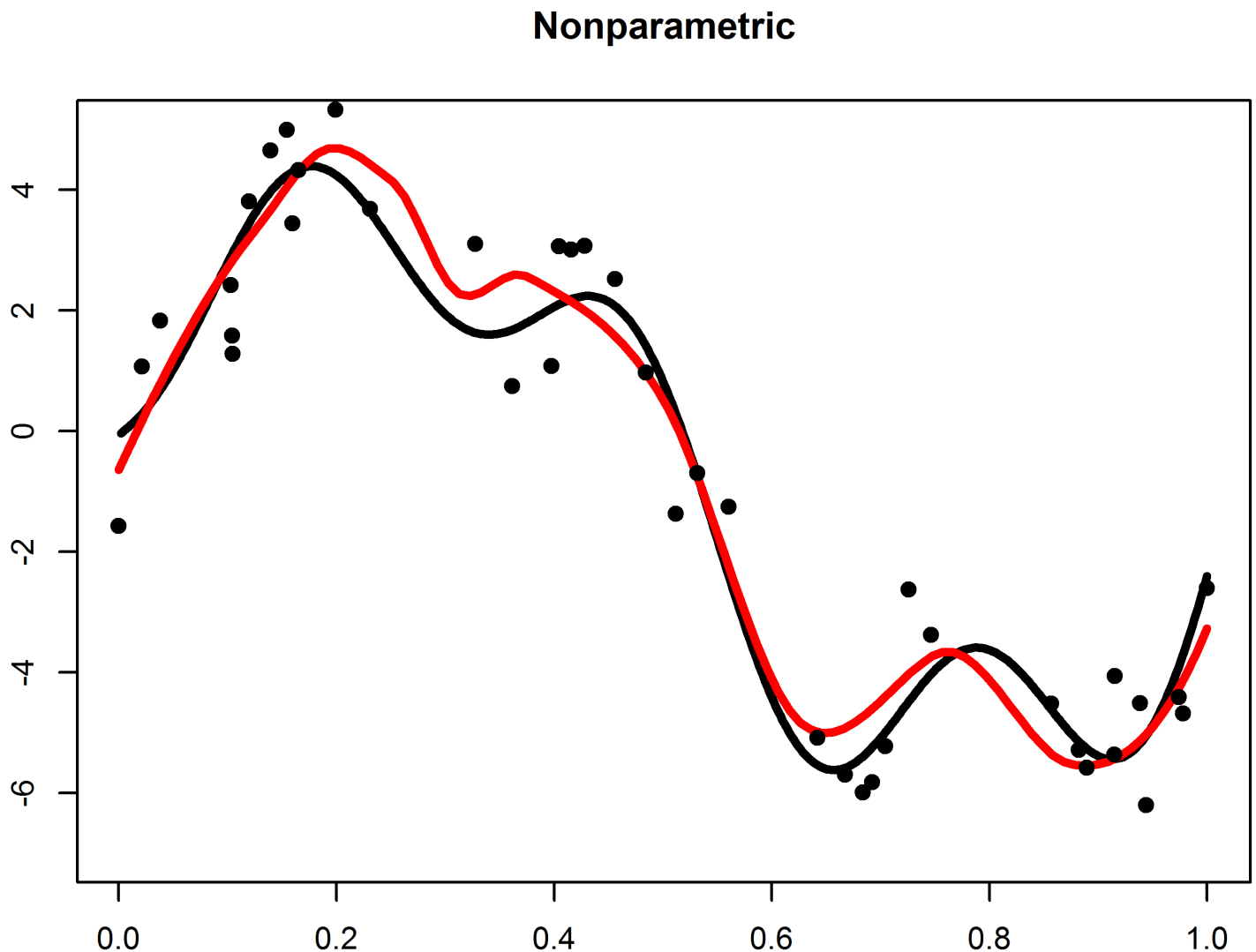
Data



Parametric fitting



Non-parametric fitting

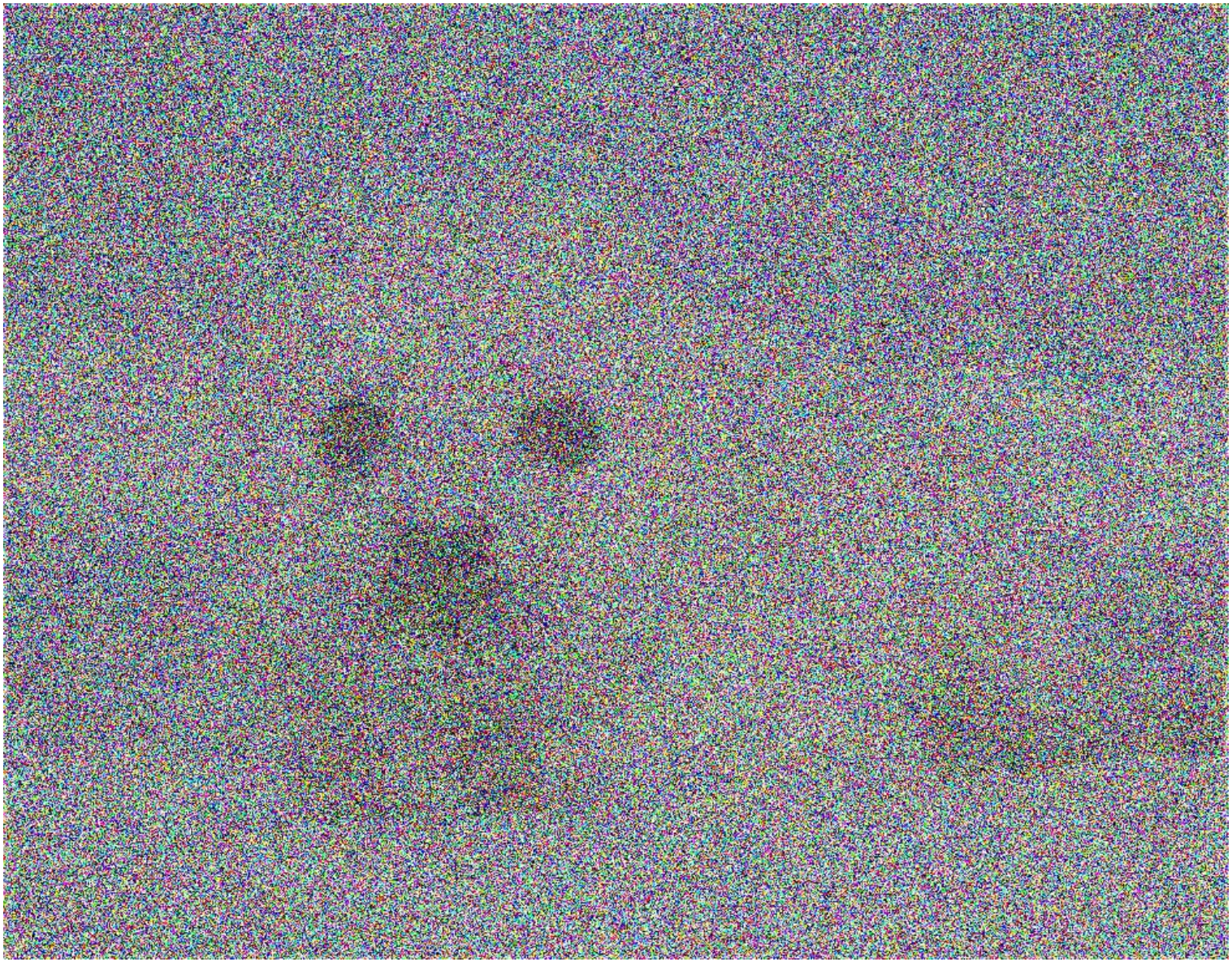


Parametric vs. Nonparametric models

Complexity of the model $\{P_\theta : \theta \in \Theta\}$:

Models:	**Parametric**	**Nonparametric**
Dimension:	Finite dimensional Θ.	Infinite dimensional Θ.
Advantages:	Easier to handle and make interpretations of the results.	Less chance for misspecifications.
	Computationally faster.	More flexible.
Disadvantages:	Without strong belief in the particular structure of the model not reliable.	Computationally and analytically challenging.
Examples:	Poisson (number of car crashes, typos in a book).	Density, regression function estimation.
	Normal distribution (grades of students, height, weight, foot-size of people).	Clustering (unknown cluster size and number).

Noisy picture



Parametric

Nonparametric

Bayesian nonparametric priors

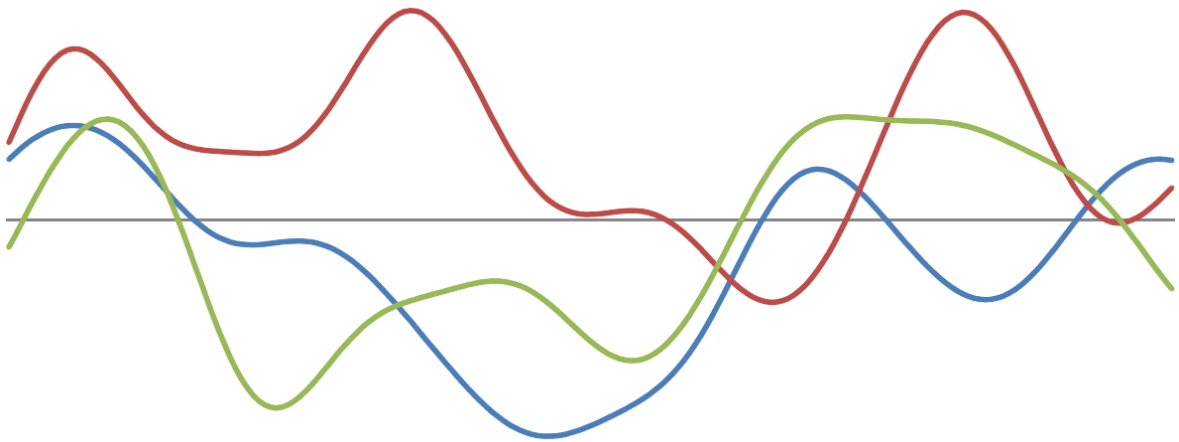
Two main categories of priors depending on parameter spaces

Bayesian nonparametric priors

Two main categories of priors depending on parameter spaces

Spaces of functions
random functions

- Continuous stochastic processes
  e.g. Gaussian processes
- Random basis expansions
- Random densities (expon.)



[Wikipedia]

Bayesian nonparametric priors

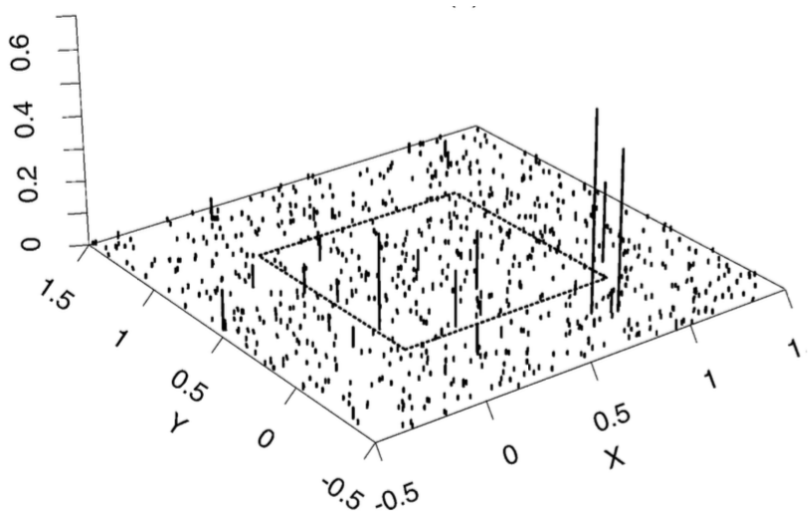
Two main categories of priors depending on parameter spaces

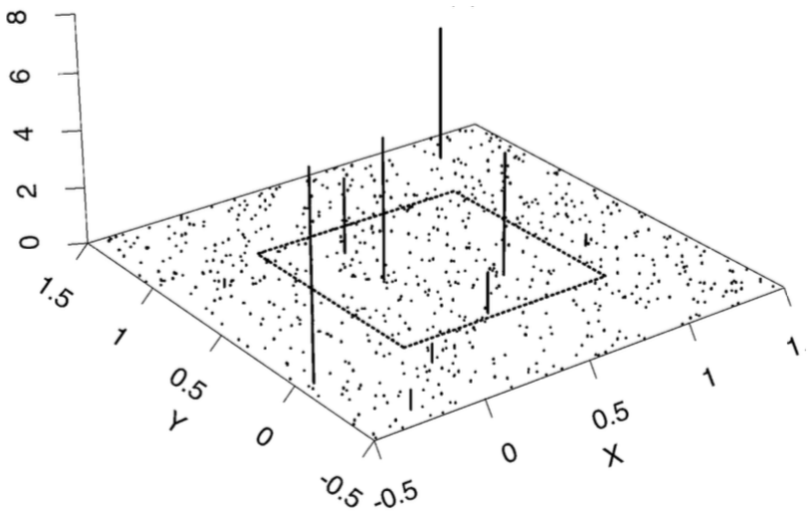
Spaces of functions
random functions

- Continuous stochastic processes
  e.g. Gaussian processes
- Random basis expansions
- Random densities (expon.)

Spaces of probability measures
random probability measures (RPM)

- Often discrete proba. measures
  Cornerstone: Dirichlet process
  We'll see others: Pitman–Yor, Normalized generalized gamma process, Normalized stable process, Gibbs-type processes, Normalized random measures, etc





[Brix, 1999]

Bayesian nonparametric priors

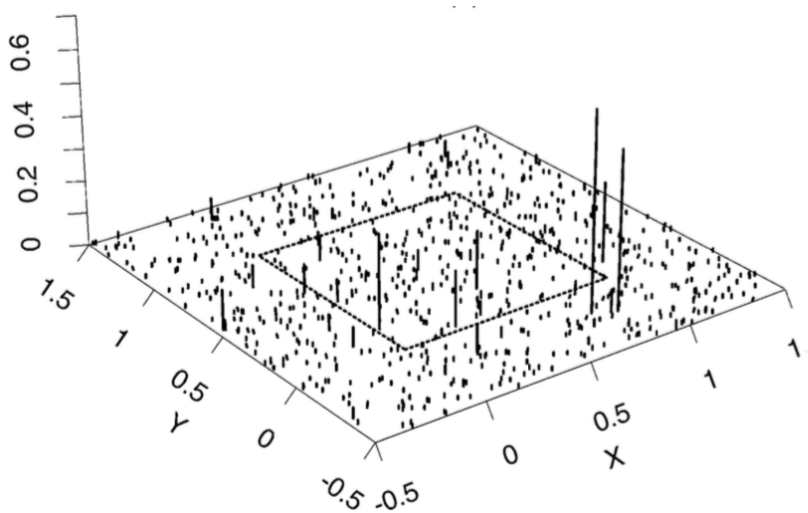
Two main categories of priors depending on parameter spaces

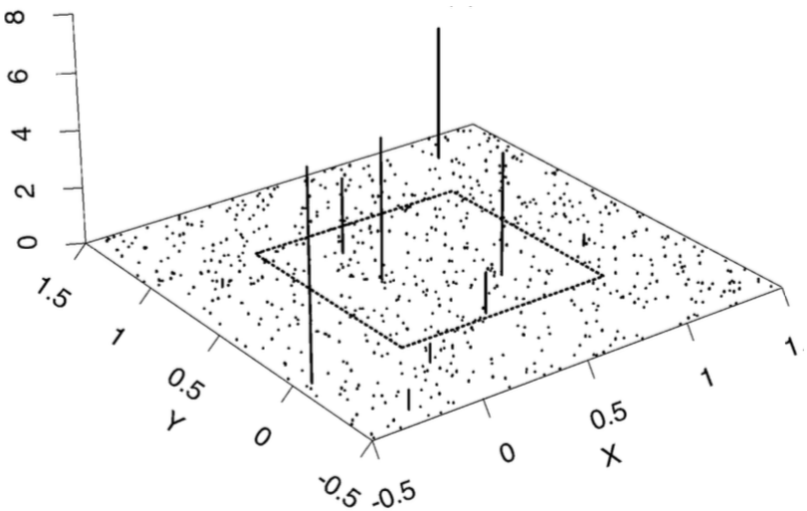
Spaces of functions
random functions

- Continuous stochastic processes
  e.g. Gaussian processes
- Random basis expansions
- Random densities (expon.)

Spaces of probability measures
random probability measures (RPM)

- Often discrete proba. measures
  Cornerstone: Dirichlet process
  We'll see others: Pitman–Yor, Normalized generalized gamma process, Normalized stable process, Gibbs-type processes, Normalized random measures, etc





[Brix, 1999]

Table of Contents

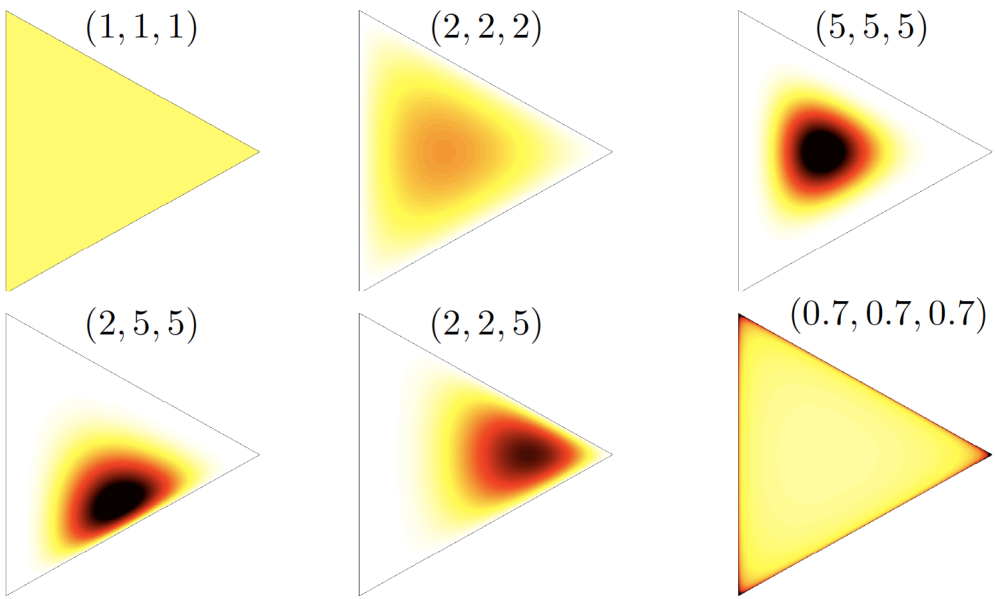
Dirichlet distribution

A *Dirichlet distribution* on a simplex Δ_K is a probability distribution with parameters $\alpha_i > 0$ and a density function

$$f(x_1, \ldots, x_K; \alpha_1, \ldots, \alpha_K) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}.$$

It is common to refer to Dirichlet distribution as $\mathrm{Dir}(\alpha_1, \ldots, \alpha_k)$.

Remark Dirichlet distribution conjugate for multinomial distribution.



[Image by Y.W. Teh]

Dirichlet process

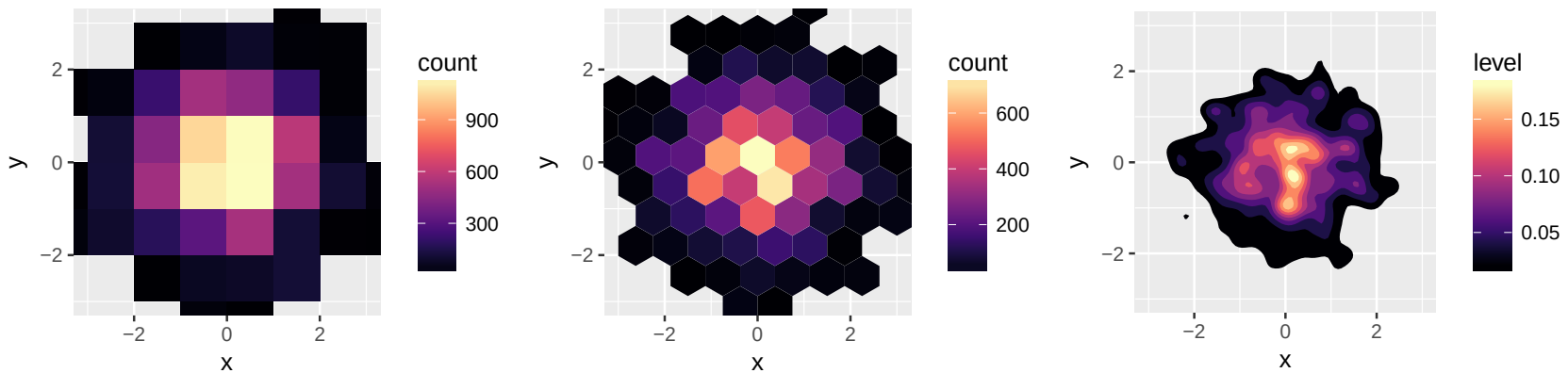
A central Bayesian nonparametric prior (Ferguson, 1973)

Definition (Dirichlet process)

A Dirichlet process on the space $\mathcal{Y}$ is a random process P such that there exist α (precision parameter) and G_0 (base/centering distribution) such that for any finite partition $\{A_1, \ldots, A_d\}$ of $\mathcal{Y}$, the random vector $(P(A_1), \ldots, P(A_d))$ is Dirichlet distributed

$$(P(A_1), \ldots, P(A_d)) \sim \text{Dir}(\alpha G_0(A_1), \ldots, \alpha G_0(A_d))$$

Notation: $P \sim DP(\alpha, G_0)$



Dirichlet process

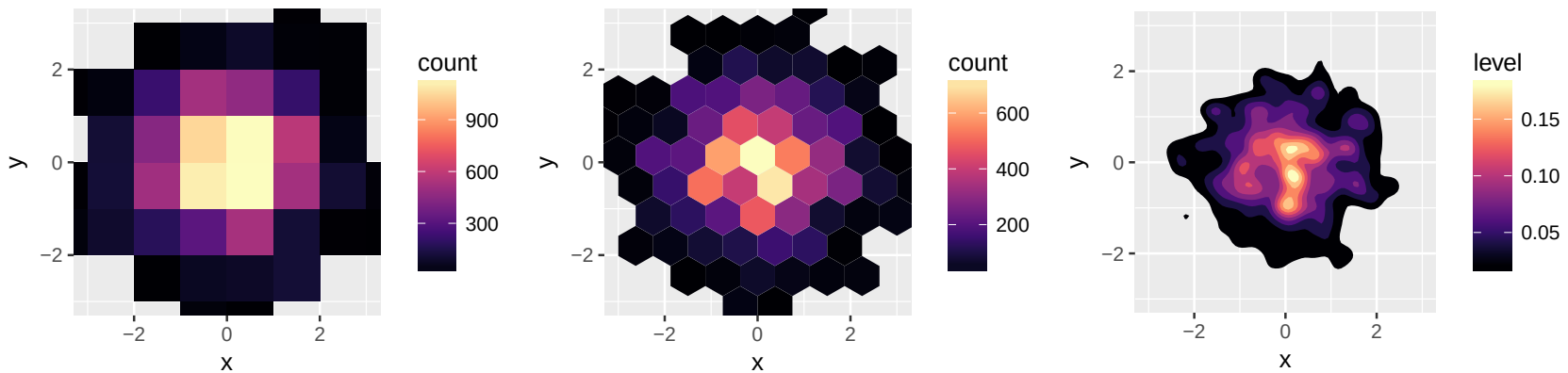
A central Bayesian nonparametric prior (Ferguson, 1973)

Definition (Dirichlet process)

A Dirichlet process on the space $\mathcal{Y}$ is a random process P such that there exist α (precision parameter) and G_0 (base/centering distribution) such that for any finite partition $\{A_1, \ldots, A_d\}$ of $\mathcal{Y}$, the random vector $(P(A_1), \ldots, P(A_d))$ is Dirichlet distributed

$$(P(A_1), \ldots, P(A_d)) \sim \text{Dir}(\alpha G_0(A_1), \ldots, \alpha G_0(A_d))$$

Notation: $P \sim DP(\alpha, G_0)$



Dirichlet process

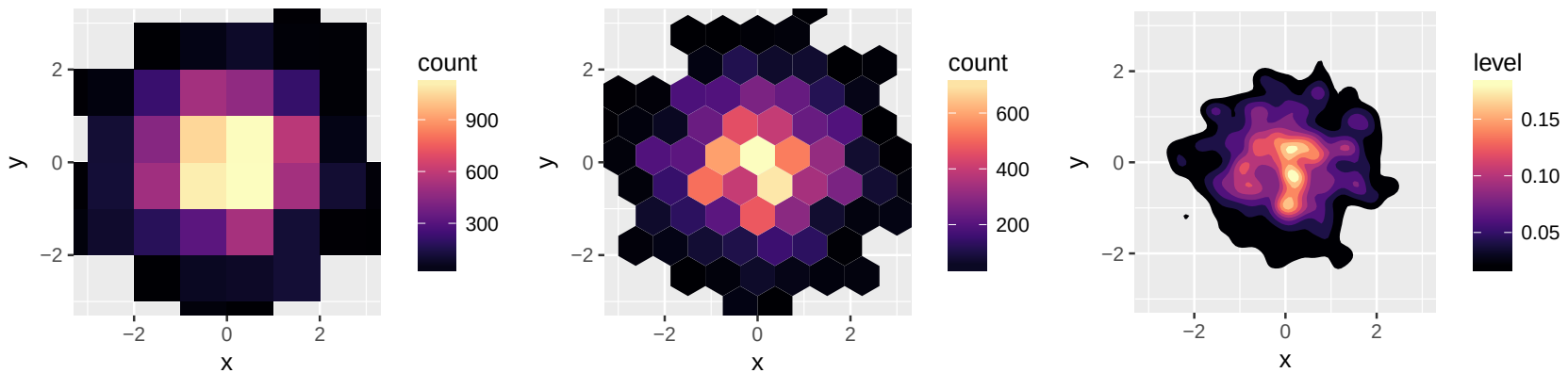
A central Bayesian nonparametric prior (Ferguson, 1973)

Definition (Dirichlet process)

A Dirichlet process on the space $\mathcal{Y}$ is a random process P such that there exist α (precision parameter) and G_0 (base/centering distribution) such that for any finite partition $\{A_1, \ldots, A_d\}$ of $\mathcal{Y}$, the random vector $(P(A_1), \ldots, P(A_d))$ is Dirichlet distributed

$$(P(A_1), \ldots, P(A_d)) \sim \text{Dir}(\alpha G_0(A_1), \ldots, \alpha G_0(A_d))$$

Notation: $P \sim DP(\alpha, G_0)$



Moments of Dirichlet process I

Proposition

Let $p \sim DP(\alpha, P_0)$ then for every measurable sets A, B we have

$$\mathbb{E}(p(A)) = P_0(A), \tag{1}$$

$$Var(p(A)) = \frac{P_0(A)(1 - P_0(A))}{1 + \alpha}, \tag{2}$$

$$\text{cov}(p(A), p(B)) = \frac{P_0(A \cap B) - P_0(A)P_0(B)}{1 + \alpha}. \tag{3}$$

Moments of Dirichlet process II

Proof

We will make use of $p(A) \sim \text{Beta}(\alpha P_0(A), \alpha(1 - P_0(A)))$. From this we obtain

$$\mathbb{E}(p(A)) = \frac{\alpha P_0(A)}{\alpha(P_0(A) + 1 - P_0(A))} = P_0(A)$$

and

$$\text{Var}(p(A)) = \frac{\alpha^2 P_0(A)(1 - P_0(A))}{\alpha^2(\alpha + 1)}.$$

We derive the covariance term in two cases, firstly taking into consideration the one with $A \cap B = \emptyset$. In that case any space Ω may be decomposed into three sets:

$$\Omega = \{A, B, (A \cup B)^c\}.$$

Using de Morgan's law the last can be written as $(A \cup B)^c = A^c \cap B^c =: C$. Therefore we may write a joint probability vector

$$\big(p(A), p(B), p(A^c \cap B^c)\big) \sim \text{Dir}\big(\alpha P_0(A), \alpha P_0(B), \alpha P_0(C)\big)$$

Moments of Dirichlet process III

and hence $\text{cov}(p(A), p(B)) = -P_0(A)P_0(B)/(1+\alpha)$. In the more general case one may decompose

$$A = (A \cap B) \cup (A \cap B^c)$$
$$B = (B \cap A) \cup (B \cap A^c),$$

so that

$$\text{cov}(P(A), P(B)) = \text{cov}(P(A \cap B) + P(A \cap B^c), P(B \cap A) + P(B \cap A^c))$$

and so forth using the linearity of covariance.

Marginalizing out the DP

Property 1 can be written equivalently as

$$\mathbb{E}(P(A)) = P_0(A) = \int P(A) dDP(P). \tag{4}$$

A Dirichlet process model can be constructed as two level sampling:

$$\begin{cases} P \sim DP(\alpha, P_0) \\ X|P \sim P, \end{cases}$$

i.e. we sample probability measure P from the Dirichlet process and then given P we sample random variables X_i.

Marginalizing out P, we obtain the marginal distribution of X:

$$X \sim P_0$$

Posterior distribution I

Let $X_1, \ldots, X_n =: X_{1:n}$ be sampled from the hierarchical model

$$\begin{cases} P \sim DP(\alpha, P_0) \\ X_{1:n}|P \overset{i.i.d.}{\sim} P, \end{cases} \tag{5}$$

This model is usually used as a building block in a larger hierarchical model, e.g. mixture models, graphs etc.

Theorem (Ferguson [1973])

The posterior of P as presented in (5) is

$$P|X_{1:n} \sim DP(\alpha P_0 + \sum_{i=1}^{n} \delta_{X_i}). \tag{6}$$

The predictive distribution of a next observation is given by

$$\mathbb{P}(X_{n+1}|X_{1:n}) = \frac{\alpha}{\alpha + n} P_0 + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{X_i}. \tag{7}$$

Posterior distribution II

The predictive (7) is also called *Polya Urn schema* or *Blackwell-MacQueen Urn Schema*.

Posterior distribution III

Proof

Property (6) can be obtained by remarking that the posterior distribution of $P(A_1), \ldots, P(A_k)$ depends on the observations only via their cell counts (it comes from *tail–free* property). Denote $N_j = \#\{1 \leq i \leq n : x_i \in A_j\}$, i.e. the number of observations in each partition of X. Then we have

$$\big(P(A_1), \ldots, P(A_k)\big)\big|X_{1:n} \stackrel{\mathrm{d}}{=} \big(P(A_1), \ldots, P(A_k)\big)\big|N_{1:k}.$$

Lets use shorthand notation: $\alpha = (\alpha_1, \ldots, \alpha_k) = (P(A_1), \ldots, P(A_k))$ and $N = (N_1, \ldots, N_k)$. Then

$$\begin{cases} N|P \sim \mathrm{Multinom}_k(P(A_1), \ldots, P(A_k)) \\ (P(A_1), \ldots, P(A_k)) \sim \mathrm{Dir}_k(\alpha P_0(A_1), \ldots, \alpha P_0(A_k)) \end{cases}$$

and hence we obtain the prior of the form

$$p(\alpha) \propto \alpha_1^{\alpha P_0(A_1)-1} \ldots \alpha_k^{\alpha P_0(A_k)-1},$$

while sampling model is

$$p(N|\alpha) \propto \alpha_1^{N_1} \ldots \alpha_k^{N_k}.$$

Posterior distribution IV

This results in the posterior of form

$$p(\alpha|N) \propto \alpha_1^{\alpha P_0(A_1)+N_1-1} \dots \alpha_k^{\alpha P_0(A_k)+N_k-1} = \mathrm{Dir}_k\left(\alpha P_0(A_1)+N_1, \dots, \alpha P_0(A_k)+N_k\right).$$

Property (7) is a result of taking the expected value of (6).

Combinatorial properties: Number of distinct values I

Assume that the base measure P_0 is non-atomic. Then with probability 1:

$$X_i \notin \{X_1, \ldots, X_{i-1}\} \Leftrightarrow X_i \sim P_0.$$

Let $D_i = \mathbb{I}(X_i \text{ is a new value})$ and lets denote $K_n = \sum_{i=1}^{n} D_i$, a number of distinct values $X_1, \ldots, X_n$ with distribution $\mathcal{L}(K_n)$.

Proposition

Random variables D_i are distributed i.i.d. with respect to Bernoulli$(\alpha/(\alpha + i - 1))$. Therefore for fixed α and for $n \to \infty$ we have:

i) $\mathbb{E}K_n \sim \alpha \log n \sim Var(K_n)$

ii) $K_n / \log(n) \xrightarrow{a.s.} \alpha$

iii) $(K_n - \mathbb{E}K_n)/sd(K_n) \to N(0, 1)$

iv) $d_{TV}\big(\mathcal{L}(K_n), Poisson(\mathbb{E}K_n)\big) = o\big(1/\log(n)\big)$ *where*

$$d_{TV}(P, Q) = sup|P(A) - Q(A)|$$

over measurable partition A

Combinatorial properties: Number of distinct values II

Proof

i) $\mathbb{E}K_n = \sum_{i=1}^{n} \frac{\alpha}{\alpha+i-1}$ and $\mathrm{Var}(K_n) = \sum_{i=1}^{n} \frac{\alpha(i-1)}{(\alpha+i-1)^2}$.

ii) Since D_i's are $\mathbb{I}$ one may use Kolmogorov law of strong numbers and

$$\sum_{i=1}^{\infty} \frac{\mathrm{Var}(D_i)}{(\log i)^2} = \sum_{i=1}^{\infty} \frac{\alpha(i-1)}{(\alpha+i-1)^2(\log i)^2} < \infty$$

by e.g. the fact that $\sum_i (1/i(\log i)^2)$ converges.

iii) By Lindeberg central limit theorem.

iv) This is implied from Chein–Stein approximation.

Combinatorial properties: Number of distinct values III

Theorem

Suppose X_i are i.i.d. such that $\mathbb{E}X_i = \mu_i$ and $VarX_i = \sigma_i^2 < \infty$. Define $Y_i = X_i - \mu_i$, $T_n = \sum_{i=1}^n Y_i$, $s_n^2 = Var(T_n) = \sum_{i=1}^n \sigma_i^2$. Then provided that

$$\forall \epsilon > 0 \quad \frac{1}{s_n^2} \sum_{i=1}^{n} \mathbb{E}\left(Y_i^2 \mathbb{I}(|Y_i| > \epsilon s_n) \right) \xrightarrow{n \to \infty} 0$$

we have $T_n / s_n \xrightarrow{d} N(0,1)$.

Combinatorial properties: Distribution of distinct values I

We have now the limits of K_n and we know its approximate distribution $\mathcal{L}(K_n)$. The **exact distribution of** K_n is:

Proposition

If P_0 is non-atomic then

$$\mathbb{P}(K_n = k) = \mathfrak{C}_n(k) n! \alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)}, \tag{8}$$

where

$$\mathfrak{C}_n(k) = \frac{1}{n!} \sum_{S \in \mathfrak{J}_n(k)} \prod_{j \in S} j \tag{9}$$

and $\mathfrak{J}_n(k) = \{S \subset \{1, \ldots, n-1\},\ |S| = n - k\}$.

Recall the definition of the **Gamma function** $\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du$.

Combinatorial properties: Distribution of distinct values II

Let us consider when we may deal with events $K_n = k$: we have two cases

$$\begin{cases} K_{n-1} = k-1 \text{ and } X_n \text{ is a new value} \\ K_{n-1} = k \text{ and } X_n \text{ is not a new value.} \end{cases}$$

This results in

$$p_n(k,\alpha) := \mathbb{P}(k_n = k|\alpha) = \frac{\alpha}{\alpha+n-1} p_{n-1}(k-1,\alpha) + \frac{n-1}{\alpha+n-1} p_{n-1}(k,\alpha). \tag{10}$$

Now let us remark that $\mathfrak{C}_n(k) = p_n(k, \alpha = 1)$. Therefore

$$\mathfrak{C}_n(k) = \frac{1}{n}\mathfrak{C}_{n-1}(k-1) + \frac{n-1}{n}\mathfrak{C}_{n-1}(k). \tag{11}$$

By induction over n: first we check case $n = 1$:

$$p_1(1,\alpha) = \mathfrak{C}_1(1)\frac{\alpha}{\alpha} = \mathfrak{C}_1(1). \tag{12}$$

Combinatorial properties: Distribution of distinct values III

To check case $n > 1$ we use (8) and then (10):

$$
\begin{aligned}
p_n(k,\alpha) &= \frac{\alpha}{\alpha+n-1}p_{n-1}(k-1,\alpha) + \frac{n-1}{\alpha+n-1}p_{n-1}(k,\alpha) \\
&= \frac{\alpha}{\alpha+n-1}\mathfrak{C}_{n-1}(k-1)(n-1)!\alpha^{k-1}\frac{\Gamma(\alpha)}{\Gamma(\alpha+n-1)} + \\
&+ \frac{n-1}{\alpha+n-1}\mathfrak{C}_{n-1}(k)(n-1)!\alpha^{k}\frac{\Gamma(\alpha)}{\Gamma(\alpha+n-1)} \\
&= \frac{\alpha^k}{\alpha+n-1}(n-1)!\frac{\Gamma(\alpha)}{\Gamma(\alpha+n-1)}n\left(\frac{1}{n}\mathfrak{C}_{n-1}(k-1) + \frac{n-1}{n}\mathfrak{C}_{n-1}(k)\right) \\
&= \mathfrak{C}_n(k)n!\alpha^k\frac{\Gamma(\alpha)}{\Gamma(\alpha+n)},
\end{aligned}
$$

which proves property (8).

Combinatorial properties: Distribution of distinct values IV

To prove (9) let use define a polynomial $A_n(s)$ as $A_n(s) = \sum_{i=1}^{\infty} \mathfrak{C}_n(k)s^k$. Then using (11) polynomial $A_n(s)$ can be written as

$$\begin{aligned} A_n(s) &= \sum_{k=1}^{\infty} \left(\frac{1}{n}\mathfrak{C}_{n-1}(k-1) + \frac{n-1}{n}\mathfrak{C}_{n-1}(k) \right) s_k \\ &= \frac{1}{n}(sA_{n-1}(s) + (n-1)A_{n-1}(s)) = \frac{s+n-1}{n}A_{n-1}(s) \\ &= \ldots = A_1(s)\prod_{j=2}^{n} \frac{s+j-1}{j} = \frac{s(s+1)\cdot\ldots\cdot(s+n-1)}{n!}. \end{aligned}$$

Last equality implies from the fact that $\mathfrak{C}_1(k) = 1\delta_{k1}$ and hence $A_1(s) = s$. Checking terms after the expansion finishes the proof of (9).

Combinatorial properties: Chinese Restaurant process I

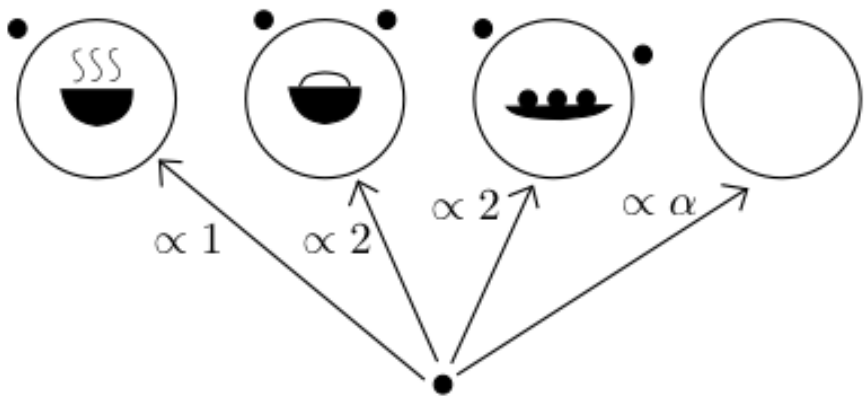
Chinese restaurant process: a culinary metaphor of the **random partition induced by the DP**. Customers join a populated table with probability $n_j/(\alpha + n)$, where n_j denotes the number of clients already sitting around the table or sit at new table with probability $\alpha/(\alpha + n)$.

Proposition

A random sample $X_{1:n}$ from a DP with precision parameter α induces a partition of $\{1, \ldots, n\}$ into k sets of sizes $n_1, \ldots, n_k$ with probability

$$p(n_1, \ldots, n_k) = p(\{n_1, \ldots, n_k\}) = \alpha^k \frac{\Gamma(\alpha)}{\Gamma(\alpha + n)} \prod_{j=1}^{k} \Gamma(n_j).$$

Combinatorial properties: Chinese Restaurant process II



Combinatorial properties: Chinese Restaurant process III

Proof

We will use the Polya urn schema slightly changed by using $n_1, \ldots, n_k$

$$\mathbb{P}(X_{n+1}|X_{1:n}) = \frac{\alpha}{\alpha + n} P_0 + \frac{1}{\alpha + n} \sum_{j=1}^{k} n_j \delta_{X_j^*}.$$

By exchangeability, the distribution of $\{n_1, \ldots, n_k\}$ does not depend on the order of the observations. Let's compute $p(n_1, \ldots, p_k)$ as the probability of one draw where the first table consists of first n_1 observations etc.
To proceed, let us use Polya urn scheme: we denote $\bar{n}_j = \sum_{i=1}^{j} n_i$ and hence $\bar{n}_k = n$, the total number of observations. We can observe the following pattern: first ball open new table, following $n_j - 1$ ones fill in that table and so forth. That quantity can be rewritten as

$$\frac{\alpha^k}{\alpha(\alpha+1)\ldots(\alpha+n-1)} \prod_{j=1}^{k} (n_j - 1)!,$$

Combinatorial properties: Chinese Restaurant process IV

where one can rewrite both terms using Gamma function $\Gamma(x) = \int_0^\infty u^{x-1} e^{-u} du$: the first term can be written as

$$\frac{\alpha^k}{\alpha(\alpha+1)\dots(\alpha+n-1)} = \frac{\Gamma(\alpha+n)}{\Gamma(\alpha)},$$

while the second one as $(n_j - 1)! = \Gamma(n_j)$.
One should remark that for ordered partitions we have

$$\bar{p}(n_1, \dots, n_k) = \frac{p(n_1, \dots, n_k)}{k!}.$$

Combinatorial properties: Ewens sampling formula I

Ewens sampling formula (ESF), presented originally by Ewens [1972], is the distribution of multiplicities $m = (m_1, \ldots, m_n)$, m_ℓ is the number of groups of size ℓ.
Also known as allelic partitions in population genetics, when there is no selective difference between types: null hypothesis in non Darwinian theory.

PROPOSITION (Ewens [1972]; Antoniak [1974])

Random variables $X_1, \ldots, X_n$ generated from a DP has multiplicity class $(m_1, \ldots, m_n)$ with probability

$$p(m_1, \ldots, m_n) = \frac{\alpha^k}{\alpha_{(n)}} \frac{n!}{\prod_{\ell=1}^n \ell^{m_\ell} m_\ell!}.$$

Notation $n_{(k)} := n(n-1) \cdot \ldots \cdot (n-k+1)$.

Combinatorial properties: Ewens sampling formula II

Proof

Two steps: 1) Compute probability of particular sequence of $X_1, \ldots, X_n$ in given class $(m_1, \ldots, m_n)$, note that all such sequences are equally likely and 2) multiply obtained quantity by the number of such sequences.

1) Consider a sequence $X_1, \ldots, X_n$ such that $X_1, \ldots, X_{m_1}$ occur each only once, then the next m_2 occur only twice and so on. This sequence has probability which may be obtained by the Polya Urn scheme in the same fashion as CRP:

$$\frac{\alpha^{m_1}(\alpha \cdot 1)^{m_2} \ldots \big(\alpha \cdot 1 \cdot \ldots \cdot (n-1)\big)^{m_n}}{\alpha_{(n)}} = \frac{\alpha^k}{\alpha_{(n)}} \prod_{\ell=1}^{n} ((\ell-1)!)^{m_\ell}.$$

2) Number of sequences $X_1, \ldots, X_n$ with frequencies $(m_1, \ldots, m_n)$ is a number of ways of putting n distinct objects into bins, so called multinomial coefficient. Since ordering of the m_ℓ bins of frequency ℓ is irrelevant, divide by $m_\ell!$:

$$\frac{1}{\prod_{l=1}^{n}(m_\ell)!} \binom{n}{1 \times \#m_1, 2 \times \#m_2, \ldots, n \times \#m_n} = \frac{n!}{\prod_{\ell=1}^{n} m_\ell!(\ell!)^{m_\ell}}$$

To finish one needs to multiply results obtained in 1) and 2).

Stick-breaking representation

The DP has almost surely discrete realizations (Sethuraman, 1994)

$$P = \sum_{j=1}^{\infty} \pi_j \delta_{\theta_j}$$

- locations $\theta_j \overset{\text{iid}}{\sim} G_0$
- weights $\pi_j = \tilde{\pi}_j \prod_{l<j}(1-\tilde{\pi}_l)$ with $\tilde{\pi}_j \overset{\text{iid}}{\sim} Beta(1, \alpha)$,

Stick-breaking representation

The DP has almost surely discrete realizations (Sethuraman, 1994)

$$P = \sum_{j=1}^{\infty} \pi_j \delta_{\theta_j}$$

- locations $\theta_j \overset{\text{iid}}{\sim} G_0$
- weights $\pi_j = \tilde{\pi}_j \prod_{l<j}(1-\tilde{\pi}_l)$ with $\tilde{\pi}_j \overset{\text{iid}}{\sim} \mathit{Beta}(1,\alpha)$,

Stick-breaking representation

The DP has almost surely discrete realizations (Sethuraman, 1994)

$$P = \sum_{j=1}^{\infty} \pi_j \delta_{\theta_j}$$

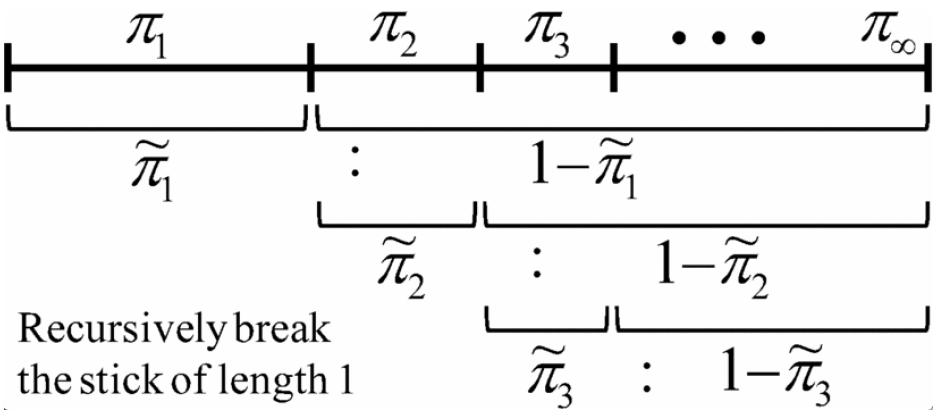
- locations $\theta_j \overset{\text{iid}}{\sim} G_0$
- weights $\pi_j = \tilde{\pi}_j \prod_{l<j}(1-\tilde{\pi}_l)$ with $\tilde{\pi}_j \overset{\text{iid}}{\sim} \mathit{Beta}(1,\alpha)$,

Stick-breaking representation

The DP has almost surely discrete realizations (Sethuraman, 1994)

$$P = \sum_{j=1}^{\infty} \pi_j \delta_{\theta_j}$$

- locations $\theta_j \overset{\text{iid}}{\sim} G_0$
- weights $\pi_j = \tilde{\pi}_j \prod_{l<j}(1-\tilde{\pi}_l)$ with $\tilde{\pi}_j \overset{\text{iid}}{\sim} \mathit{Beta}(1,\alpha)$,

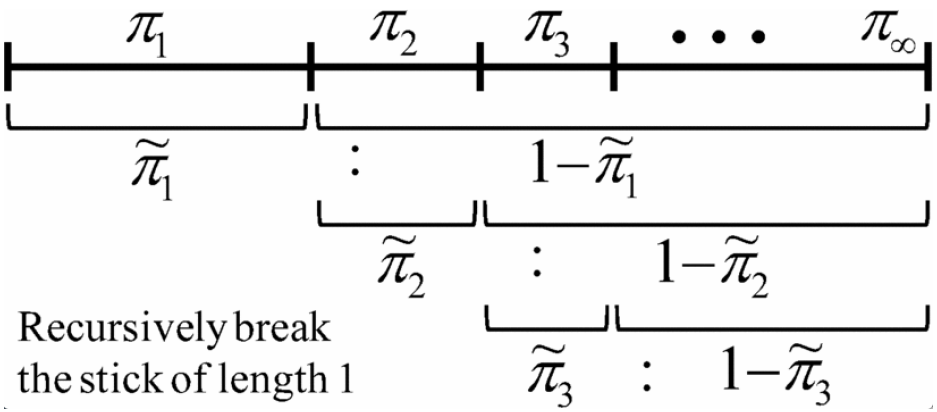


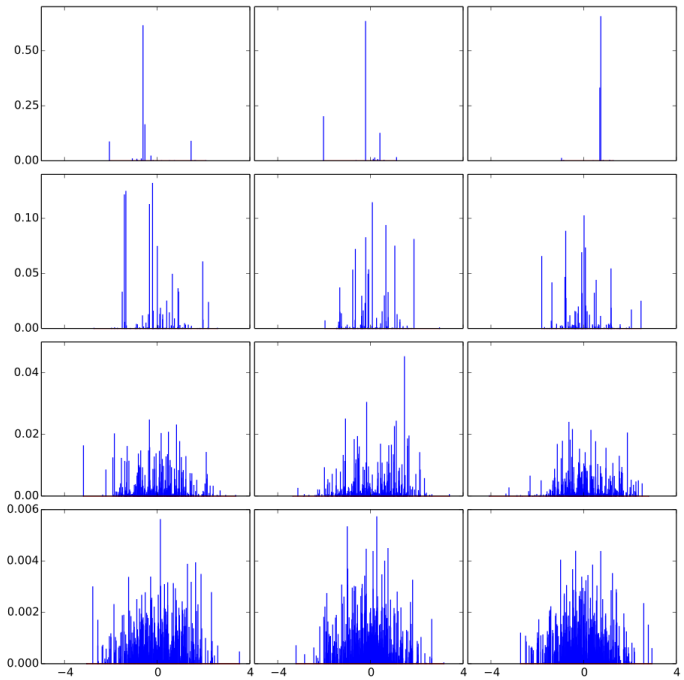
Stick-breaking representation

The DP has almost surely discrete realizations (Sethuraman, 1994)

$$P = \sum_{j=1}^{\infty} \pi_j \delta_{\theta_j}$$

- locations $\theta_j \overset{\text{iid}}{\sim} G_0$
- weights $\pi_j = \tilde{\pi}_j \prod_{l<j}(1-\tilde{\pi}_l)$ with $\tilde{\pi}_j \overset{\text{iid}}{\sim} \mathit{Beta}(1,\alpha)$,





Stick-breaking representation I

Theorem (Sethuraman [1994])

If $V_1, V_2, \ldots \overset{i.i.d.}{\sim} Be(1, \alpha)$ *and* $\phi_1, \phi_2, \ldots \overset{i.i.d.}{\sim} P_0$ *are i.i.d. variables, then define* $p_1 = V_1$ *and*

$$p_j = V_j \prod_{1 \leq l \leq j} (1 - V_l)$$

then

$$P = \sum_{i=1}^{\infty} p_i \delta_{\phi_i} \sim DP(\alpha, P_0).$$

Stick-breaking representation II

Lemma

For independent $\phi \sim P_0$ and $V \sim Be(1, \alpha)$ the DP is the only solution of the distributional equation

$$P \stackrel{\mathrm{d}}{=} V\delta_\phi + (1 - V)P, \tag{13}$$

where $P \sim DP(\alpha, P_0)$.

Stick-breaking representation III

Proof

1) The weights $(p_1, p_2, \ldots)$ need to form a probability vector. The leftover mass at stage j is

$$1 - \left(\sum_{i=1}^{j} p_i \right) = \prod_{i=1}^{j} (1 - V_i) =: R_j.$$

One may notice that R_j is decreasing and for every j we have $R_j \in [0, 1]$, hence we obtain almost sure convergence which is equivalent with convergence in mean. Therefore

$$\mathbb{E} R_j = \mathbb{E} \prod_j (1 - V_j) = \prod_j \mathbb{E}(1 - V_j) = \left(\frac{\alpha}{\alpha + 1} \right)^j \to 0.$$

So $(p_1, \ldots)$ is a probability vector almost surely and P is a probability measure almost surely.

Stick-breaking representation IV

2) Now one may write

$$P = p_1\delta_{\phi_1} + \sum_{j=2}^{\infty} p_j\delta_{\phi_j} = V_1\delta_{\phi_1} + (1 - V_1)\sum_{j=1}^{\infty} \tilde{p}_j\delta_{\tilde{\phi}_j},$$

where $\tilde{p}_j = \frac{p_{j+1}}{1-V_1} = V_{j+1}\prod_{l=2}^{j}(1 - V_l)$ and $\tilde{\phi}_j = \phi_{j+1}$, then $(\tilde{p}_j)$ and $(\tilde{\phi}_j)$ satisfy the same distributional definitions as (p_j) and (ϕ_j), hence $\tilde{P} \overset{\mathrm{d}}{=} P$ and so P is solution of the Lemma equation (13) whose only solution is the DP.

DP as a normalized Gamma process I

The DP can be obtained by normalizing a Gamma process. It is a generic way to obtain independently distributed probability measures from almost surely finite random measures. Let us investigate for the case $\mathcal{Y} = \mathbb{R}$.

Definition

Gamma process on $\mathbb{R}_+$ is a process $(S(u) : u \geq 0)$ with independent increments satisfying

$$\forall u_1 : 0 \leq u_1 \leq u_2 : \quad S(u_2) - S(u_1) \overset{\perp}{\sim} Ga(u_2 - u_1, 1).$$

This ensures that the process has non-decreasing right continuous sample path $u \mapsto S(u)$.

Theorem

For every $\alpha > 0$ and for every cumulative distribution function G, a random cumulative distribution function such that

$$F(t) = \frac{S(\alpha G(t))}{S(\alpha)}$$

is the distribution of a $DP(\alpha, G)$.

DP as a normalized Gamma process II

Proof

For any set of t_i satisfying $-\infty = t_0 < t_1 < \ldots < t_k = \infty$ we have

$$S(\alpha G(t_i)) - S(\alpha G(t_{i-1})) \sim Ga\big(\alpha G(t_i) - \alpha G(t_{i-1}), 1\big).$$

Use property that if $Y_i \overset{ind}{\sim} Ga(\alpha_i, 1)$ then
$(Y_1, \ldots, Y_n)/\sum_i Y_i \sim \text{Dir}_n(\alpha_1, \ldots, \alpha_n)$ to obtain

$$\big(F(t_1)-F(t_0), \ldots, F(t_k)-F(t_{k-1})\big) \sim \text{Dir}_k\big(\alpha G(t_1)-\alpha G(t_0), \ldots, \alpha G(t_k)-\alpha G(t_{k-1})\big).$$

Hence the definition of DP holds for every partition in intervals. These form a measure determining class, so that the definition holds for every partition in general.

Definition via the Polya Urn Scheme

A Polya sequence with parameter αP_0 is a sequence of random variables $X_1, \ldots, X_n$ whose joint distribution satisfies

$$X_1 \sim P_0, \;\; X_{n+1}|X_1, \ldots, X_n \sim \frac{\alpha}{\alpha + n} P_0 + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{X_i}. \tag{14}$$

Theorem

If $X_1, X_2, \ldots$ is a Polya sequence then exists random probability measure P such that $X_i|P \overset{i.i.d.}{\sim} P$ and $P \sim DP(\alpha, P_0)$.

Proof

We can consider Polya sequence as an outcome of Polya urn, we see that it is exchangeable. By de Finetti theorem exists such probability measure P such that $X_i|P \overset{i.i.d.}{\sim} P$. So far we have proved existence of the DP and know that DP generates a Polya sequence. Since the RPM given by de Finetti's theorem is unique this proves that $P \sim DP(\alpha, P_0)$.

Definition via the Polya Urn Scheme

A Polya sequence with parameter αP_0 is a sequence of random variables $X_1, \ldots, X_n$ whose joint distribution satisfies

$$X_1 \sim P_0, \;\; X_{n+1}|X_1, \ldots, X_n \sim \frac{\alpha}{\alpha + n} P_0 + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{X_i}. \tag{14}$$

Theorem

If $X_1, X_2, \ldots$ is a Polya sequence then exists random probability measure P such that $X_i|P \overset{i.i.d.}{\sim} P$ and $P \sim DP(\alpha, P_0)$.

Proof

We can consider Polya sequence as an outcome of Polya urn, we see that it is exchangeable. By de Finetti theorem exists such probability measure P such that $X_i|P \overset{i.i.d.}{\sim} P$. So far we have proved existence of the DP and know that DP generates a Polya sequence. Since the RPM given by de Finetti's theorem is unique this proves that $P \sim DP(\alpha, P_0)$.

Definition via the Polya Urn Scheme

A Polya sequence with parameter αP_0 is a sequence of random variables $X_1, \ldots, X_n$ whose joint distribution satisfies

$$X_1 \sim P_0, \;\; X_{n+1}|X_1, \ldots, X_n \sim \frac{\alpha}{\alpha + n} P_0 + \frac{1}{\alpha + n} \sum_{i=1}^{n} \delta_{X_i}. \tag{14}$$

Theorem

If $X_1, X_2, \ldots$ is a Polya sequence then exists random probability measure P such that $X_i|P \overset{i.i.d.}{\sim} P$ and $P \sim DP(\alpha, P_0)$.

Proof

We can consider Polya sequence as an outcome of Polya urn, we see that it is exchangeable. By de Finetti theorem exists such probability measure P such that $X_i|P \overset{i.i.d.}{\sim} P$. So far we have proved existence of the DP and know that DP generates a Polya sequence. Since the RPM given by de Finetti's theorem is unique this proves that $P \sim DP(\alpha, P_0)$.

Table of Contents

A parametric approach

Mixture Model with K components

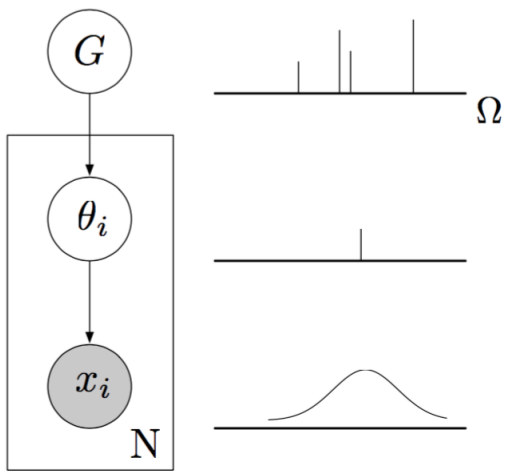
$$G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k}$$

δ_{ϕ_k} is a point mass at ϕ_k.
G is to be understood as a K-faceted dice. The mixture density is:

$$p(X|\pi, \phi) = \sum_{k=1}^{K} \pi_k p(x|\phi_k)$$

Then

$$\begin{aligned} \theta_i &\sim G \\ x_i &\sim p(x|\theta_i) \end{aligned}$$



A Bayesian parametric approach

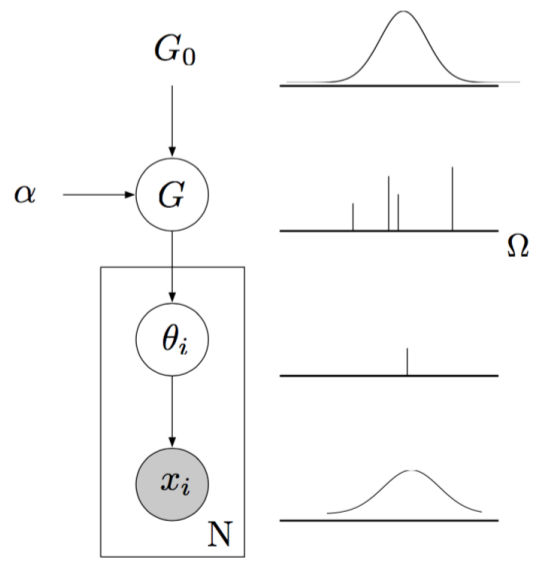
Bayesian Mixture Models with K components

We need a distribution over the probability measure (aka dice) G, that is a distribution over weights or classes $\pi = (\pi_1, \ldots, \pi_K)$ and over mean and covariance (for 2-dimensional data) $\phi_k = (\mu_k, \Sigma_k)$

- $\pi \sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)$
- $(\mu_k, \Sigma_k) \sim \text{Normal} \times \text{Inverse-Wishart}$

This makes $G = \sum_{k=1}^{K} \pi_k \delta_{\phi_k}$ a random dice

$$
\begin{aligned}
\phi_k &\sim G_0 \\
\pi &\sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K) \\
G &= \sum_{i=1}^{K} \pi_k \delta_{\phi_k} \\
\theta_i &\sim G \\
x_i &\sim p(x|\theta_i)
\end{aligned}
$$



Choosing K

There are several options for choosing K

- Model selection with information criteria: AIC, BIC, or cross-validation, etc
- Hierarchical model, with a prior on K
- Be nonparametric, and let K get large... possibly infinite.

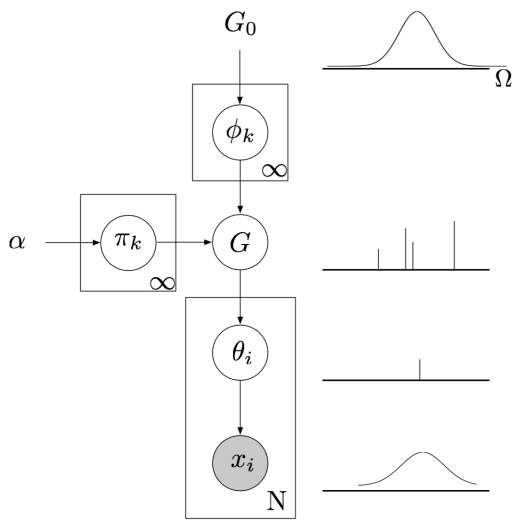
A Bayesian nonparametric approach

Bayesian nonparametric Mixture Models

We now move to G being an infinite sum $G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$

We need a distribution over this infinite dice G, that is exactly what the **Dirichlet process** does. It is parameterized by the precision parameter α and the base measure G_0.

- $\pi = (\pi_1, \pi_2, \ldots) \sim \text{GEM}(\alpha)$
- $\phi_k \sim G_0$



Posterior sampling

Markov chain Monte Carlo (MCMC) methods

- **Marginal methods**: marginalizing over the posterior DP P, and sampling using the posterior Pólya urn scheme (easy in conjugate case) Neal [2000]
- **Conditional methods**: sampling a finite but sufficient number of parameters
  - ϵ-DP Muliere and Tardella [1998], ϵ-PY Arbel et al. [2018b]
  - Blocked Gibbs sampler Ishwaran and James [2001]
  - Slice sampler Walker [2007]
  - Retrospective sampler Papaspiliopoulos and Roberts [2008]
  - Ferguson and Klass algorithm Ferguson and Klass [1972]; Arbel and Prünster [2017]
- **Variational approximations**

Sampling from the posterior distribution

Packages that do precisely that!

- **DPpackage**: Jara et al. [2011], implemented in Fortran
- **BNPdensity**: Barrios et al. [2013], extension to flexible classes of priors (normalized random measures)
- **BNPmix** (ongoing): Arbel et al. [2018a], Dirichlet process mixtures, fast implementation with **Rcpp** package. Can be installed by:

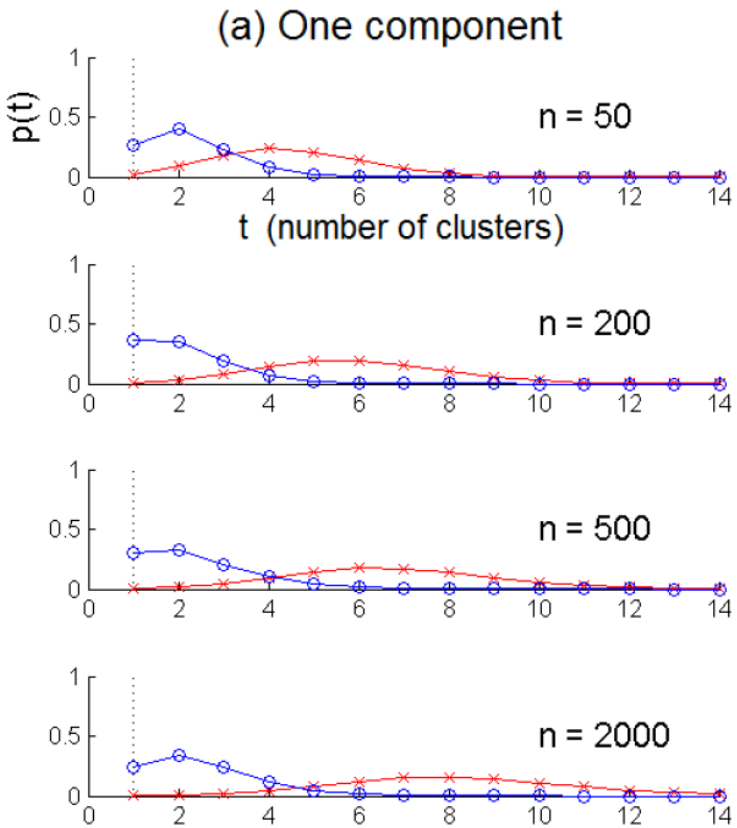
```
library(devtools)
install_github("rcorradin/BNPmix")
```

Warning on interpretation of K_n I

Consider a simple DP mixture model with

- Gaussian base measure,
- Gaussian kernel,
- where data are sampled iid from some distribution.

Then the **posterior on K_n is inconsistent** [Miller and Harrison, 2013].



Warning on interpretation of K_n II

From Miller and Harrison [2013] (here K_n is denoted T_n):

Theorem 4.1. *If $X_1, X_2, \ldots \in \mathbb{R}$ are i.i.d. from any distribution with $\mathbb{E}|X_i| < \infty$, then with probability 1, under the standard normal DPM with $\alpha = 1$ as defined above, $p(T_n = 1 \mid X_{1:n})$ does not converge to 1 as $n \to \infty$.*

Theorem 5.1. *If $X_1, X_2, \ldots \sim \mathcal{N}(0, 1)$ i.i.d. then*

$$p(T_n = 1 \mid X_{1:n}) \xrightarrow{\Pr} 0 \quad \text{as } n \to \infty$$

under the standard normal DPM with concentration parameter $\alpha = 1$.

But there is some hope...

Bayesian decision theory

From decision theory: a Bayes estimator minimizes a posterior expected loss.

$$\hat{a}_L = \arg \inf_{a \in A} \mathbb{E}_{\pi(\theta)}[L_a(\theta)].$$

Examples with Euclidean parameter spaces:

- L^2, squared loss $\longrightarrow$ posterior mean
- L^1, absolute loss $\longrightarrow$ posterior median
- $0-1$ loss $\longrightarrow$ mode a posteriori (MAP)

Bayesian decision theory

From decision theory: a Bayes estimator minimizes a posterior expected loss.

$$\hat{a}_L = \arg \inf_{a \in A} \mathbb{E}_{\pi(\theta)}[L_a(\theta)].$$

Examples with Euclidean parameter spaces:

- L^2, squared loss $\longrightarrow$ posterior mean
- L^1, absolute loss $\longrightarrow$ posterior median
- $0-1$ loss $\longrightarrow$ mode a posteriori (MAP)

Deriving an optimal clustering

The posterior expected loss of clustering c', denoted by $L(c')$, is obtained by **averaging the loss with respect to posterior weight**

$$L(c') = \sum_{c \in \mathcal{A}_n} L(c, c') p(c|\mathbf{x}),$$

and the decision is taken by choosing the best

$$\hat{c} = \arg \min_{c' \in \mathcal{A}_n} \sum_{c \in \mathcal{A}_n} L(c, c') p(c|\mathbf{x})$$

Several losses have been considered:

- 0-1 loss [Rajkowski, 2016],
- Binder loss [Dahl, 2006],
- Variation of information [Wade and Ghahramani, 2018].

Deriving an optimal clustering

The posterior expected loss of clustering c', denoted by $L(c')$, is obtained by **averaging the loss with respect to posterior weight**

$$L(c') = \sum_{c \in \mathcal{A}_n} L(c, c') p(c|\mathbf{x}),$$

and the decision is taken by choosing the best

$$\hat{c} = \arg \min_{c' \in \mathcal{A}_n} \sum_{c \in \mathcal{A}_n} L(c, c') p(c|\mathbf{x})$$

Several losses have been considered:

- 0-1 loss [Rajkowski, 2016],
- Binder loss [Dahl, 2006],
- Variation of information [Wade and Ghahramani, 2018].

Simplest loss: L_{0-1}

$$L_{0-1}(c') = \sum_{c \in \mathcal{A}_n} L_{0-1}(c, c')p(c|\mathbf{x}) = \sum_{c \in \mathcal{A}_n,\, c \neq c'} p(c|\mathbf{x}),$$
$$= 1 - p(c'|\mathbf{x})$$

which is to say that the expected loss of c' is **all the posterior mass except that of c'.** So that it is easily minimized at the value c' which has **maximum** posterior weight:

$$\hat{c} = \arg \min_{c' \in \mathcal{A}_n} L_{0-1}(c') = \arg \max_{c' \in \mathcal{A}_n} p(c'|\mathbf{x}) := MAP.$$

Negative results by Rajkowski [2016] show that the **mode a posteriori (MAP) is inconsistent**.

Simplest loss: L_{0-1}

$$L_{0-1}(c') = \sum_{c \in \mathcal{A}_n} L_{0-1}(c, c') p(c|\mathbf{x}) = \sum_{c \in \mathcal{A}_n,\, c \neq c'} p(c|\mathbf{x}),$$
$$= 1 - p(c'|\mathbf{x})$$

which is to say that the expected loss of c' is **all the posterior mass except that of c'.** So that it is easily minimized at the value c' which has **maximum** posterior weight:

$$\hat{c} = \arg \min_{c' \in \mathcal{A}_n} L_{0-1}(c') = \arg \max_{c' \in \mathcal{A}_n} p(c'|\mathbf{x}) := MAP.$$

Negative results by Rajkowski [2016] show that the **mode a posteriori (MAP) is inconsistent**.

Simplest loss: L_{0-1}

$$L_{0-1}(c') = \sum_{c \in \mathcal{A}_n} L_{0-1}(c, c') p(c|\mathbf{x}) = \sum_{c \in \mathcal{A}_n,\, c \neq c'} p(c|\mathbf{x}),$$
$$= 1 - p(c'|\mathbf{x})$$

which is to say that the expected loss of c' is **all the posterior mass except that of c'.** So that it is easily minimized at the value c' which has **maximum** posterior weight:

$$\hat{c} = \arg \min_{c' \in \mathcal{A}_n} L_{0-1}(c') = \arg \max_{c' \in \mathcal{A}_n} p(c'|\mathbf{x}) := MAP.$$

Negative results by Rajkowski [2016] show that the **mode a posteriori (MAP) is inconsistent**.

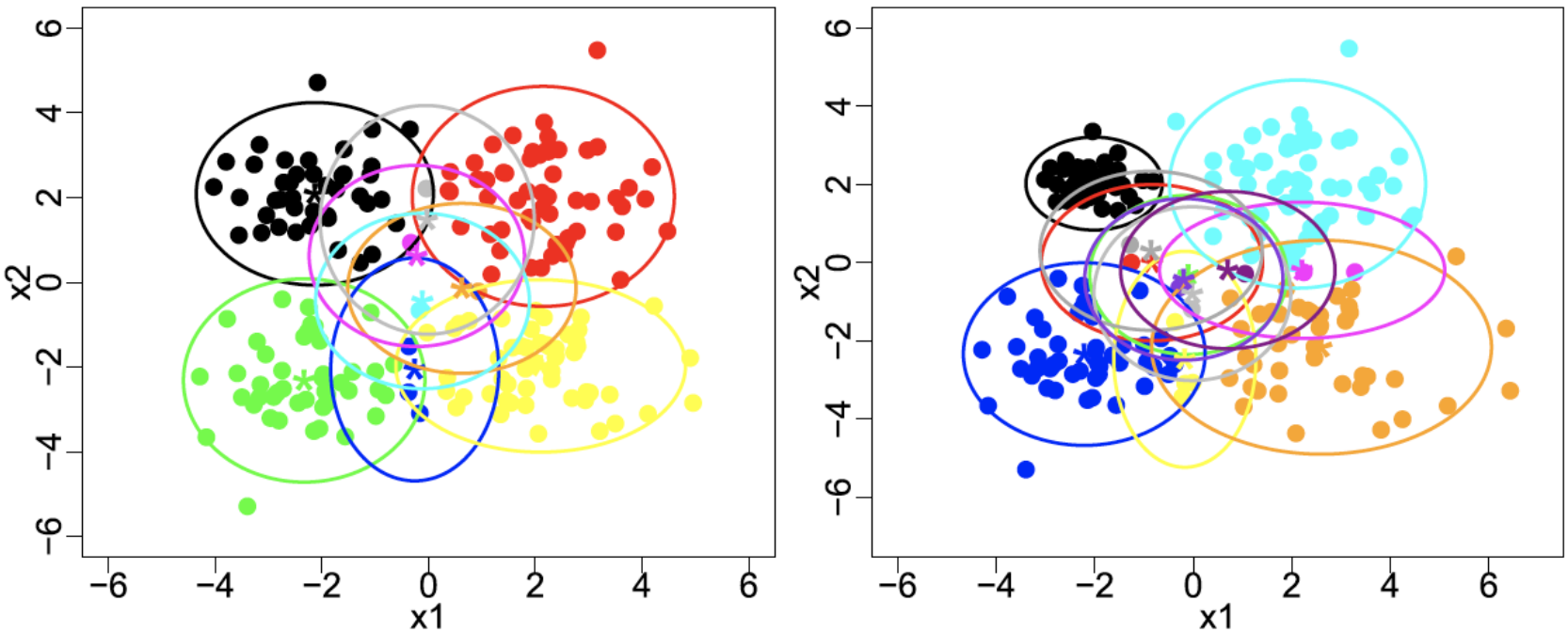
Variation of information

Variation of information (VI) by Meilă [2007] for cluster comparison. From information theory, compares information in two clusterings with information shared between the two clusterings:

$$\mathrm{VI}(c, \widehat{c}) = \mathrm{H}(c) + \mathrm{H}(\widehat{c}) - 2\mathcal{I}(c, \widehat{c})$$

Variation of information

Wade and Ghahramani [2018] compare Binder and VI:

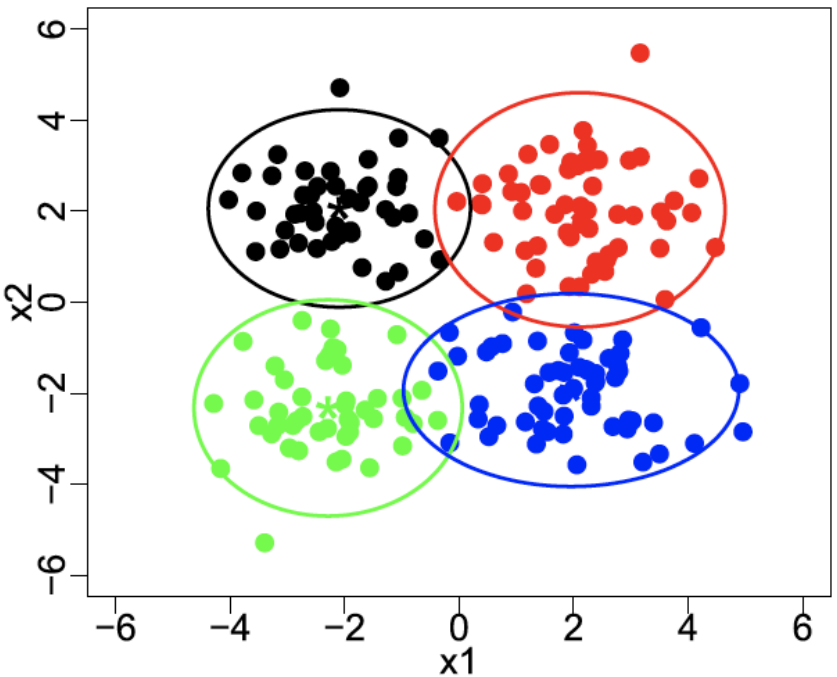


(a) Ex 1 Binder's: 9 clusters

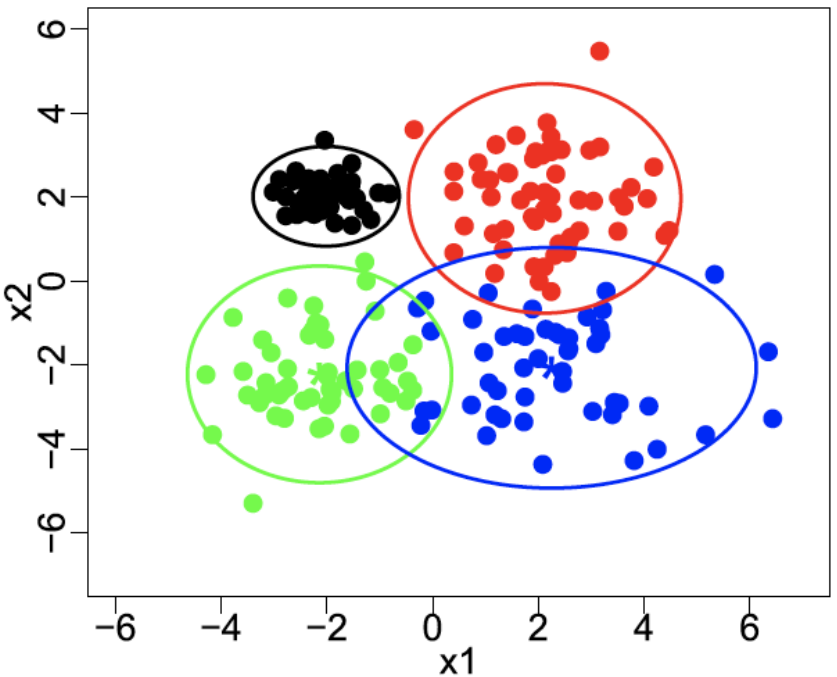
(b) Ex 2 Binder's: 12 clusters

Variation of information

Wade and Ghahramani [2018] compare Binder and VI:



(c) Ex 1 VI: 4 clusters



(d) Ex 2 VI: 4 clusters

Variation of information

Wade and Ghahramani [2018] provide **credible balls** around the estimated clustering, based on Hasse diagram:

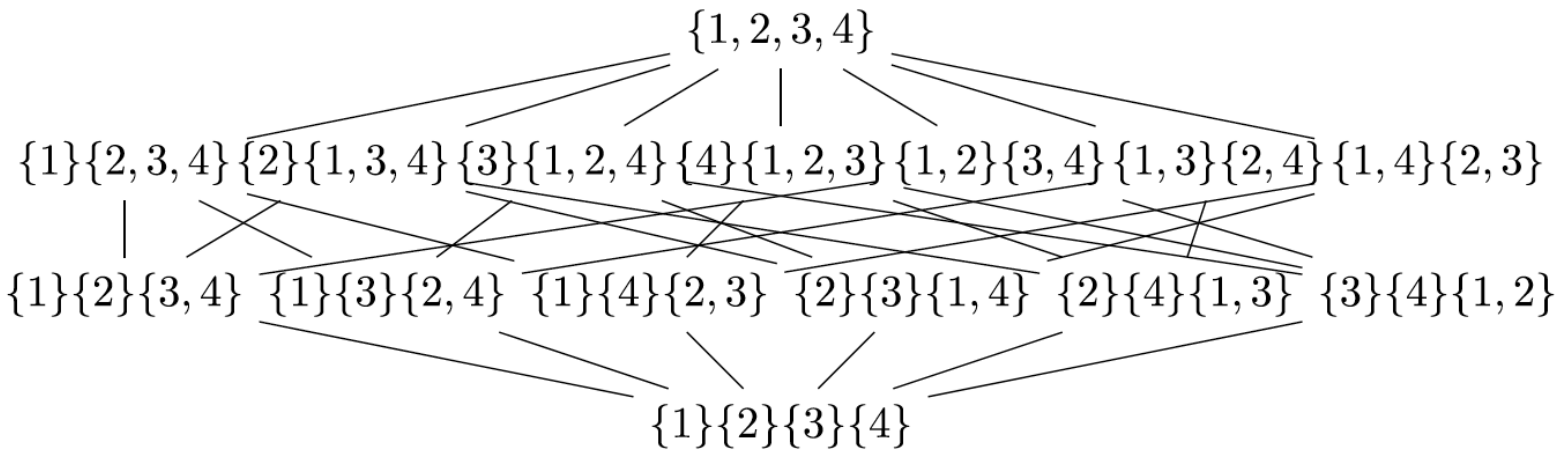


Figure 1: Hasse diagram for the lattice of partitions with a sample of size $N = 4$. A line is drawn from $\mathbf{c}$ up to $\widehat{\mathbf{c}}$ when $\mathbf{c}$ is covered by $\widehat{\mathbf{c}}$.

Table of Contents

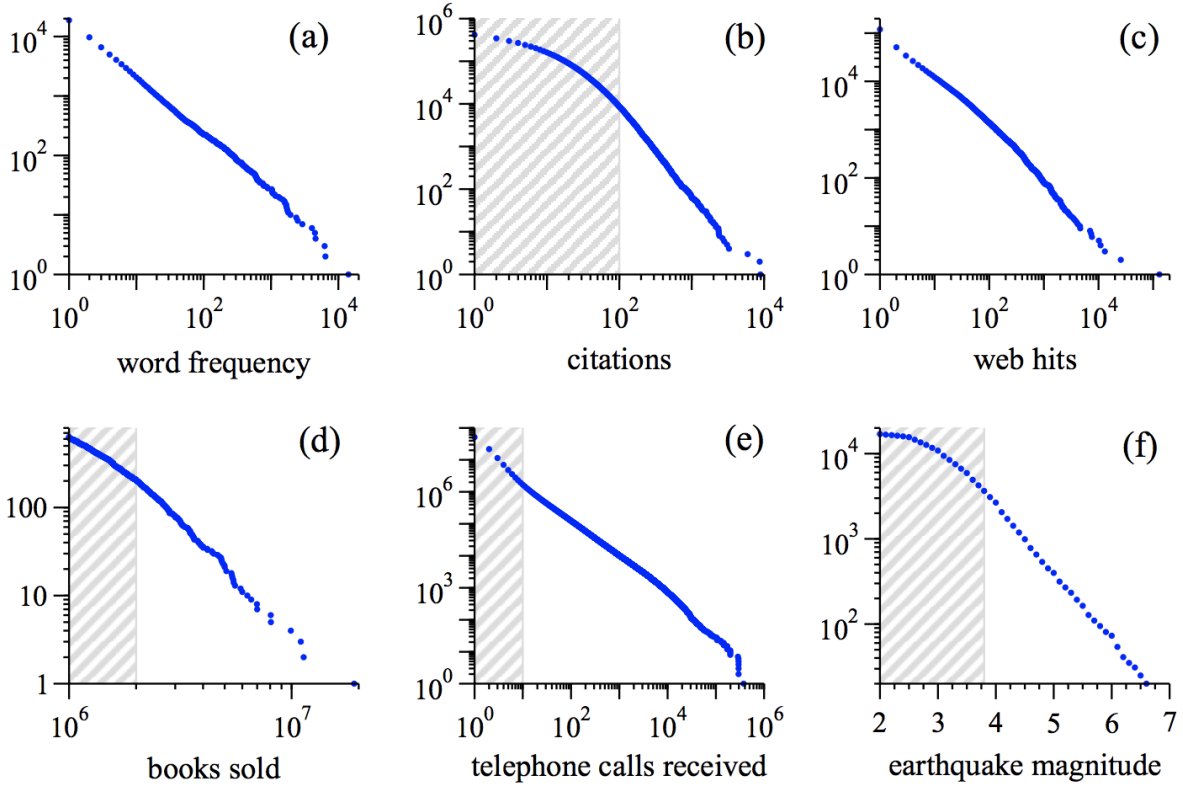
Need for a power-law for K_n

Newman [2005]; Clauset et al. [2009] show that "*Power-law distributions occur in many situations of scientific interest and have significant consequences for our understanding of natural and man-made phenomena*".

Hence the need to depart from $K_n \sim \alpha \log n$ induced by a Dirichlet process.

Need for a power-law for K_n

Newman [2005]; Clauset et al. [2009] show that "*Power-law distributions occur in many situations of scientific interest and have significant consequences for our understanding of natural and man-made phenomena*".



[Image from Newman [2005]]

Hence the need to depart from $K_n \sim \alpha \log n$ induced by a Dirichlet process.

Chinese restaurant process

Consider discrete data $X_1, \dots X_n | P \overset{\text{iid}}{\sim} P$, and $P \sim \mathcal{Q}$
Features $k_n \leq n$ unique values $X_1^*, \dots, X_{k_n}^*$ with resp. frequencies $n_1, \dots, n_{k_n}$

Discrete random probability measures are characterized by predictive distr.

Dirichlet process by Ferguson (1973): $P \sim DP(\alpha, G_0)$

$$\mathbb{P}[X_{n+1} \in \cdot \,|\, X_1, \dots X_n] = \frac{\alpha}{\alpha + n} G_0(.) + \frac{1}{\alpha + n} \sum_{j=1}^{k_n} n_j \delta_{X_j^*}(.)$$

Log rate for number of clusters $k_n \asymp \alpha \log n$

Product form exchangeable partition probability function

$$p(n_1, \dots, n_{k_n}) = \alpha^{k_n} \frac{\Gamma(\alpha)}{\Gamma(\alpha + k_n)} \prod_{j=1}^{k_n} (n_j - 1)!$$

Chinese restaurant process

Consider discrete data $X_1, \dots X_n | P \overset{\text{iid}}{\sim} P$, and $P \sim \mathcal{Q}$
Features $k_n \leq n$ unique values $X_1^*, \dots, X_{k_n}^*$ with resp. frequencies $n_1, \dots, n_{k_n}$

Discrete random probability measures are characterized by predictive distr.

Pitman–Yor process by Pitman & Yor (1997): $P \sim PY(\sigma, \alpha, G_0)$, $\sigma \in (0, 1)$

$$\mathbb{P}[X_{n+1} \in \cdot \,|\, X_1, \dots X_n] = \frac{\alpha + \sigma k_n}{\alpha + n} G_0(.) + \frac{1}{\alpha + n} \sum_{j=1}^{k_n} (n_j - \sigma) \delta_{X_j^*}(.)$$

Power law rate for number of clusters $k_n \asymp S n^{\sigma}$

Product form exchangeable partition probability function

$$p(n_1, \dots, n_{k_n}) = \frac{\prod_{i=1}^{k_n - 1} (\alpha + i\sigma)}{(\alpha + 1)_{(n-1)}} \prod_{j=1}^{k_n} (1 - \sigma)_{(n_j - 1)}$$

Chinese restaurant process

Consider discrete data $X_1, \dots X_n | P \overset{\text{iid}}{\sim} P$, and $P \sim \mathcal{Q}$
Features $k_n \leq n$ unique values $X_1^*, \dots, X_{k_n}^*$ with resp. frequencies $n_1, \dots, n_{k_n}$

Discrete random probability measures are characterized by predictive distr.

Gibbs-type processes by Pitman (2003): $P \sim Gibbs(\sigma, (V_{n,k})_{n,k}, G_0)$, $\sigma < 1$

$$\mathbb{P}[X_{n+1} \in \cdot \,|\, X_1, \dots X_n] = \frac{V_{n+1,k_n+1}}{V_{n,k_n}} G_0(.) + \frac{V_{n+1,k_n}}{V_{n,k_n}} \sum_{j=1}^{k_n} (n_j - \sigma) \delta_{X_j^*}(.)$$

Rate for number of clusters $k_n \asymp$
- K random variable a.s. finite if $\sigma < 0$
- $\alpha \log n$ if $\sigma = 0$
- Sn^σ if $\sigma \in (0,1)$, (S random variable).

Product form exchangeable partition probability function

$$p(n_1, \dots, n_{k_n}) = V_{n,k_n} \prod_{j=1}^{k_n} (1-\sigma)_{(n_j - 1)}$$

Beyond the DP from predictive function viewpoint

A discrete random probability measure P can be classified in 3 main categories according to $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n]$

1) $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n] = f(n, \text{model parameters})$
$\iff$ depends on n but not on k_n and $(n_1, \ldots, n_{k_n})$
$\iff$ Dirichlet process (Ferguson, 1973);

2) $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n] = f(n, k_n, \text{model parameters})$
$\iff$ depends on n and k_n but not on $(n_1, \ldots, n_{k_n})$
$\iff$ Gibbs-type prior (Pitman, 2003);

3) $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n] = f(n, k_n, (n_1, \ldots, n_{k_n}), \text{model parameters})$
$\iff$ depends on n, k_n and $(n_1, \ldots, n_{k_n})$
$\iff$ tractability issues

Beyond the DP from predictive function viewpoint

A discrete random probability measure P can be classified in 3 main categories according to $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n]$

1) $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n] = f(n, \text{model parameters})$
$\iff$ depends on n but not on k_n and $(n_1, \ldots, n_{k_n})$
$\iff$ Dirichlet process (Ferguson, 1973);

2) $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n] = f(n, k_n, \text{model parameters})$
$\iff$ depends on n and k_n but not on $(n_1, \ldots, n_{k_n})$
$\iff$ Gibbs-type prior (Pitman, 2003);

3) $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n] = f(n, k_n, (n_1, \ldots, n_{k_n}), \text{model parameters})$
$\iff$ depends on n, k_n and $(n_1, \ldots, n_{k_n})$
$\iff$ tractability issues

Beyond the DP from predictive function viewpoint

A discrete random probability measure P can be classified in 3 main categories according to $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n]$

1) $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n] = f(n, \text{model parameters})$
$\iff$ depends on n but not on k_n and $(n_1, \ldots, n_{k_n})$
$\iff$ Dirichlet process (Ferguson, 1973);

2) $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n] = f(n, k_n, \text{model parameters})$
$\iff$ depends on n and k_n but not on $(n_1, \ldots, n_{k_n})$
$\iff$ Gibbs-type prior (Pitman, 2003);

3) $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n] = f(n, k_n, (n_1, \ldots, n_{k_n}), \text{model parameters})$
$\iff$ depends on n, k_n and $(n_1, \ldots, n_{k_n})$
$\iff$ tractability issues

Beyond the DP from predictive function viewpoint

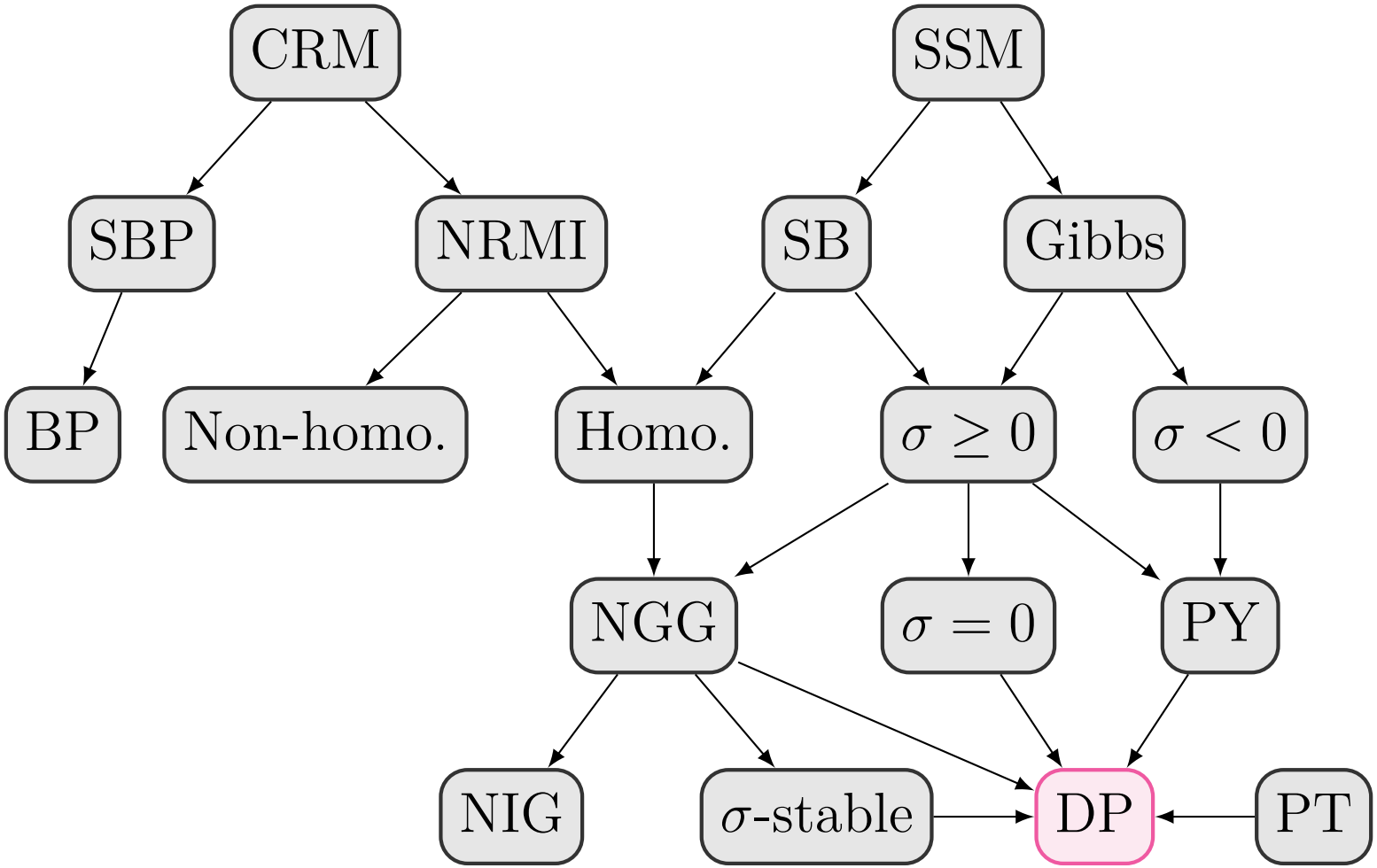
A discrete random probability measure P can be classified in 3 main categories according to $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n]$

1) $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n] = f(n, \text{model parameters})$
$\iff$ depends on n but not on k_n and $(n_1, \ldots, n_{k_n})$
$\iff$ Dirichlet process (Ferguson, 1973);

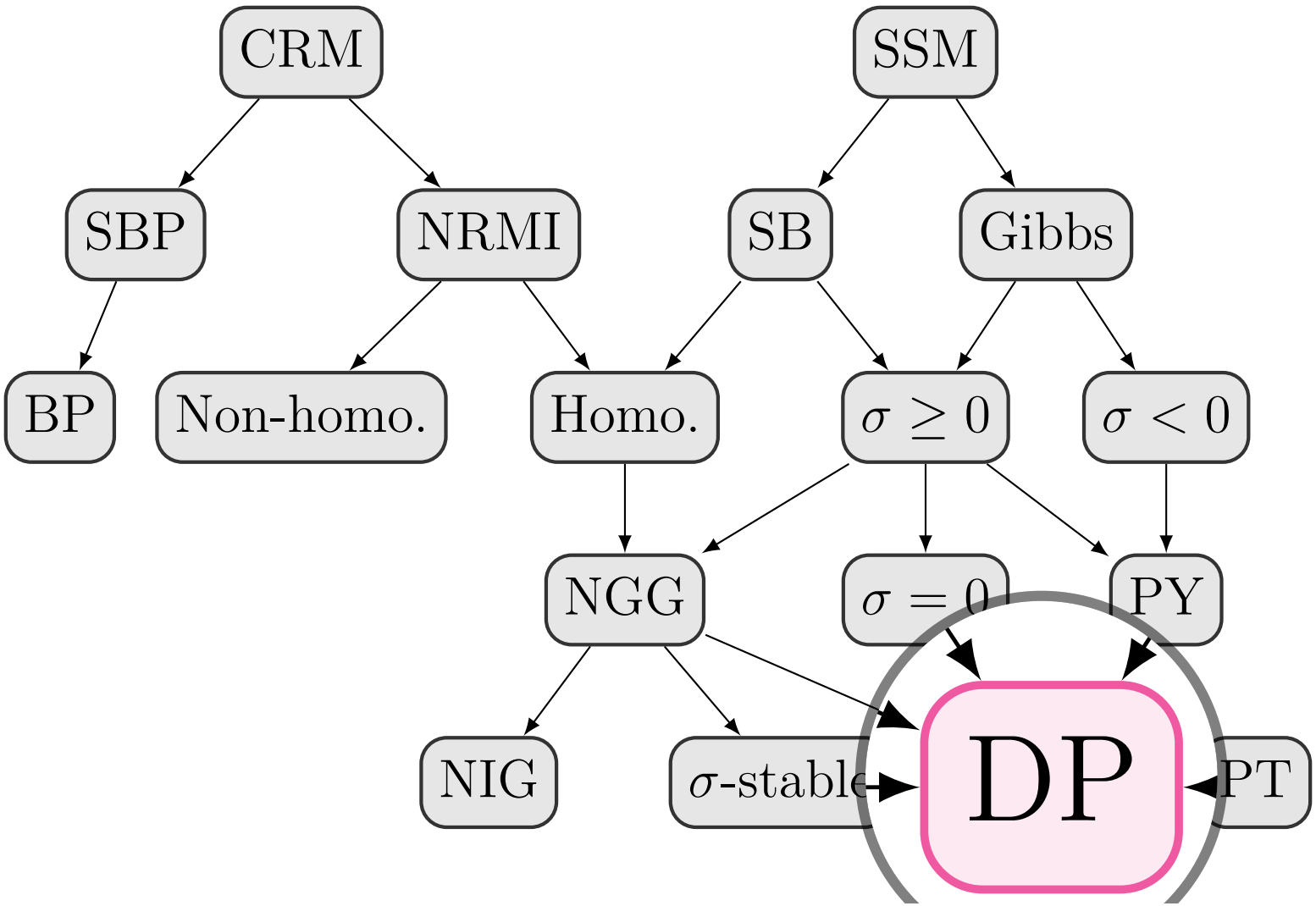
2) $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n] = f(n, k_n, \text{model parameters})$
$\iff$ depends on n and k_n but not on $(n_1, \ldots, n_{k_n})$
$\iff$ Gibbs-type prior (Pitman, 2003);

3) $\mathbb{P}[X_{n+1} \text{ is "new"} \mid \boldsymbol{X}_n] = f(n, k_n, (n_1, \ldots, n_{k_n}), \text{model parameters})$
$\iff$ depends on n, k_n and $(n_1, \ldots, n_{k_n})$
$\iff$ tractability issues

Tree of discrete random probability measures



Tree of discrete random probability measures



Pitman–Yor process I

PROPOSITION (Pitman Sampling formula)

The multiplicities $(m_1, \ldots, m_n)$ *in* $X_1, \ldots, X_n | P \overset{i.i.d.}{\sim} P$, $P \sim PY(\sigma, \alpha P_0)$ *have distribution*

$$p(m_1, \ldots, m_n) = \frac{n!}{(1+\alpha)_{(n-1)}} (\alpha+\sigma)\cdot\ldots\cdot(\alpha+(K_n-1)\sigma) \prod_{\ell=1}^{n} \frac{1}{m_\ell!} \left(\frac{(1-\sigma)_{(\ell-1)}}{\ell!} \right)^{m_\ell}$$

Proof

Same technique as for the DP ESF.

Pitman–Yor process II

Proposition

Power Law and σ-diversity

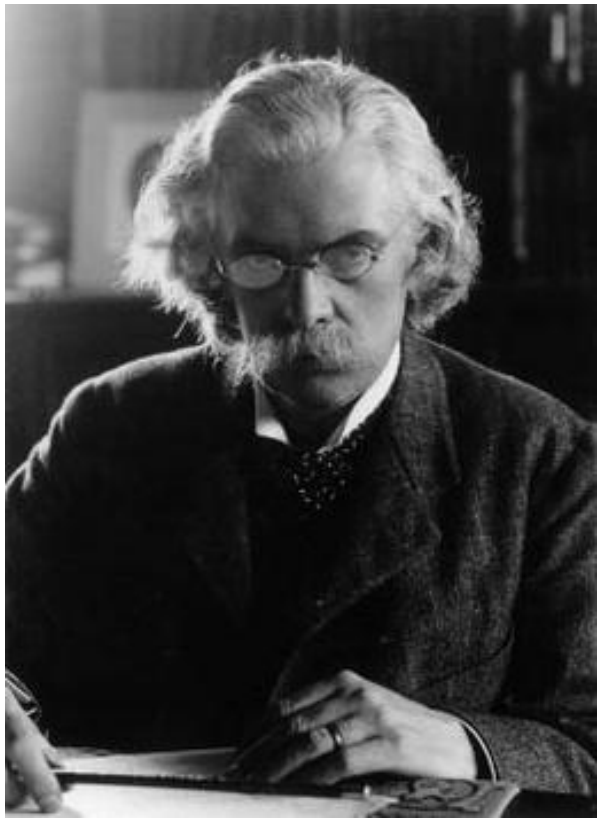
For $\sigma > 0$ we have the almost sure convergence

$$n^{-\sigma} K_n \to S_{\sigma,\alpha},$$

where $S_{\sigma,\alpha}$ is called σ-diversity of the PY, whose density is a polynomially tilted **Mittag–Leffler density** *(ML):*

$$g_{\sigma,\alpha}(x) \propto x^{\alpha/\sigma} g_\alpha(x),$$

and g_α is ML density.



[Image: Wikipedia]

Pitman–Yor process III

Theorem

Stick breaking representation for PY process

If $V_j \overset{ind}{\sim} Be(1-\sigma, \alpha + j\sigma)$ *and* $p_1 = V_1,\ p_j = V_j \prod_{l<j}(1-V_l)$ *and further we have* $\phi_j \overset{i.i.d.}{\sim} P_0$ *then*

$$P = \sum_{j=1}^{\infty} p_j \delta_{\phi_j} \sim PY(\sigma, \alpha P_0).$$

Pitman–Yor process IV

Proposition (Moments of PY)

If $P \sim PY(\sigma, \alpha P_0)$, then for every measurable sets A, B we have

1) $\mathbb{E}(P(A)) = P_0(A)$,
2) $\mathbb{E}(P(A)P(B)) = (1-\sigma)/(1+\alpha)P_0(A \cap B) + (\alpha+\sigma)/(1+\alpha)P_0(A)P_0(B)$,
3) $\mathrm{cov}(P(A), P(B)) = (1-\sigma)/(1+\alpha)\big(P_0(A \cap B) - P_0(A)P_0(B)\big)$.

Remark

As before P_0 is the mean measure, while σ lowers dependance with respect to DP.

Pitman–Yor process V

Proof

1) We use stick–breaking representation:

$$\mathbb{E}P(A) = \sum_j \mathbb{E}p_j \mathbb{E}\delta_{\phi_j} = \sum_j \mathbb{E}(p_j) P_0(A) = P_0(A)\mathbb{E}(\sum_j p_j) = P_0(A).$$

2) Let $X_1, X_2 | P \overset{i.i.d.}{\sim} P$, then

$$\mathbb{E}(P(A)P(B)) = \mathbb{P}(X_1 \in A, X_2 \in B) = \mathbb{P}(X_1 \in A)\mathbb{P}(X_2 \in B | X_1 \in A).$$

Lets investigate two terms above: from 1) we know that $\mathbb{P}(X_1 \in A) = P_0(A)$. We know the predictive of PY:

$$X_2 | X_1 \sim \frac{\alpha + \sigma}{\alpha + 1} P_0 + \frac{1 - \sigma}{\alpha + 1} \delta_{X_1},$$

and hence

$$\mathbb{P}(X_2 \in B | X_1 \in A) = \frac{\alpha + \sigma}{\alpha + 1} P_0(B) + \frac{1 - \sigma}{\alpha + 1} P_{0A}(B),$$

when we used notation $P_{0A}(B) = P_0(B|A) = P_0(A \cap B)/P_0(A)$ for a conditional measure.

3) It is straightforward combination of 1) and 2).

Pitman–Yor process VI

Unlike the DP, PY is not conjugate under incoming independent samples. However, the posterior can be explicited.

Theorem (Posterior distribution of PY)

If $P \sim PY(\sigma, \alpha P_0)$ then the posterior of P based on observations $X_{1:n}|P \overset{i.i.d.}{\sim} P$ has the distribution of the random probability measure

$$(1 - q_n)P_n + q_n \sum_{j=1}^{K_n} p_j^* \delta_{X_j^*},$$

where $X_{1:n}^*$ are the K_n distinct values in $X_{1:n}$, frequencies are refered to as $n_1, \ldots, n_{K_n}$ and

- $q_n \sim Beta(n - K_n\sigma, \alpha + K_n\sigma)$,
- $(p_1^*, \ldots, p_{K_n}^*) \sim Dir_{K_n}(n_1 - \sigma, \ldots, n_{K_n} - \sigma)$,
- $P_n \sim PY(\sigma, (\alpha + \sigma K_n)P_0)$.

Impact of the stability parameter σ

Prior distribution of the number of clusters k_n

- α controls the location (as for the DP)
- σ controls the flatness (or variability)

Impact of the stability parameter σ

Prior distribution of the number of clusters k_n

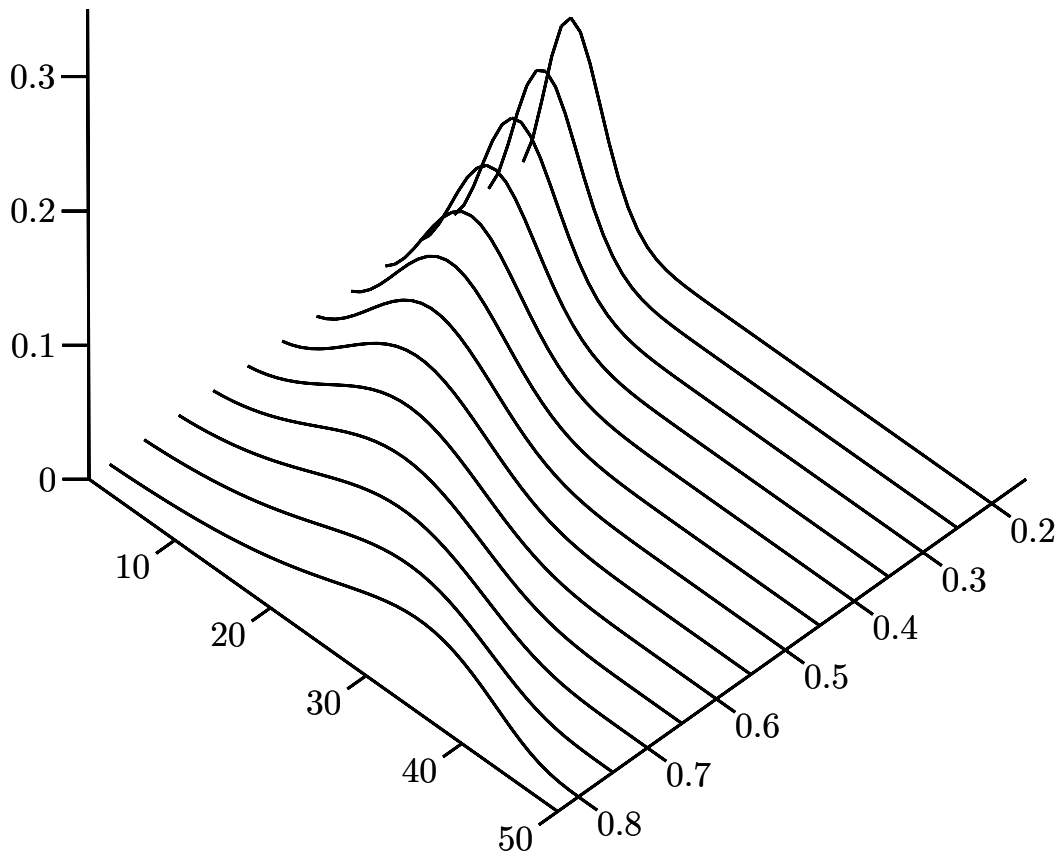
- α controls the location (as for the DP)
- σ controls the flatness (or variability)

Impact of the stability parameter σ

Prior distribution of the number of clusters k_n

- α controls the location (as for the DP)
- σ controls the flatness (or variability)

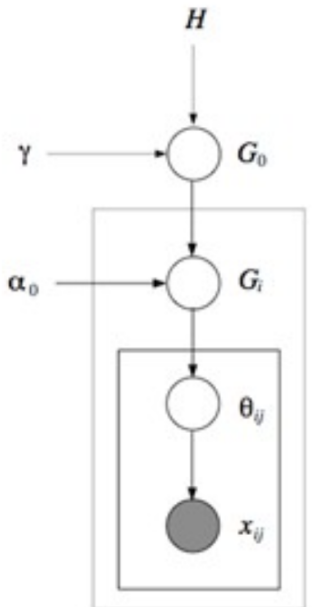
Example with $n = 50, \alpha = 1$ and $\sigma = 0.2, 0.3, \ldots, 0.8$

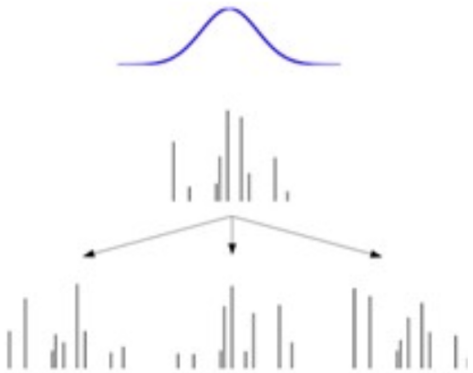


[Image by De Blasi et al. [2015]]

Hierarchical Dirichlet process Teh et al. [2006]

A nonparametric version of **Latent dirichlet allocation** [Blei et al., 2003]





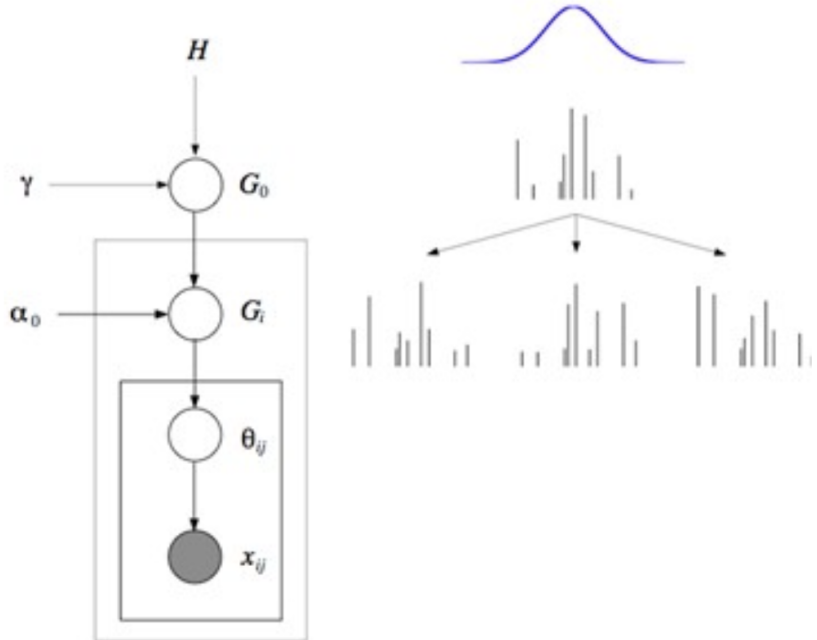
$$
\begin{aligned}
G_0|\gamma, H &\sim DP(\gamma H) \\
G_i|\alpha, G_0 &\sim DP(\alpha_0 G_0) \\
\theta_{ij}|G_i &\sim G_i \\
x_{ij}|\theta_{ij} &\sim F(x_{ij}|\theta_{ij})
\end{aligned}
$$

[Image by M. Jordan]

Associated partition distr. called Chinese Restaurant Franchise.

Hierarchical Dirichlet process Teh et al. [2006]

A nonparametric version of **Latent dirichlet allocation** [Blei et al., 2003]



$$
\begin{aligned}
G_0|\gamma, H &\sim DP(\gamma H) \\
G_i|\alpha, G_0 &\sim DP(\alpha_0 G_0) \\
\theta_{ij}|G_i &\sim G_i \\
x_{ij}|\theta_{ij} &\sim F(x_{ij}|\theta_{ij})
\end{aligned}
$$

[Image by M. Jordan]

Associated partition distr. called Chinese Restaurant Franchise.

Indian Buffet process Ghahramani and Griffiths [2006]

Feature allocation model: observations may share several features.

Generative model is as follows

- first customer samples Poisson(γ) dishes
- second customer chooses every dish of first customer *wp* 1/2, plus Poisson($\gamma/2$) new dishes
- ...
- ith step: K dishes have been sampled, each by $n_1, \ldots, n_K$ customers; ith customer chooses jth dish *wp* n_j/i, plus Poisson(γ/i) new dishes.

Log growth for K_n:

$$K_n \sim \text{Poisson}(\gamma \log n).$$

Indian Buffet process Ghahramani and Griffiths [2006]

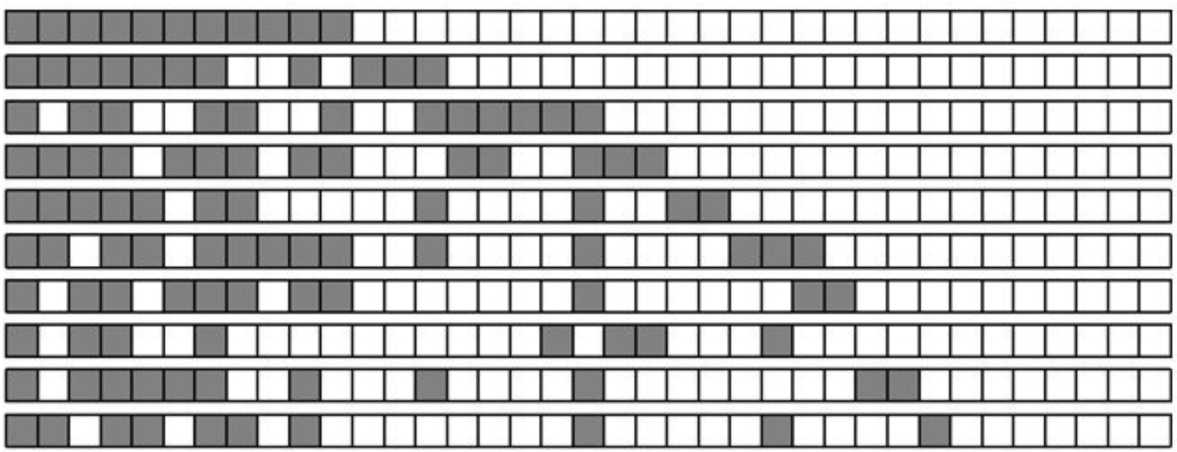
Feature allocation model: observations may share several features.

Generative model is as follows

- first customer samples Poisson(γ) dishes
- second customer chooses every dish of first customer *wp* $1/2$, plus Poisson($\gamma/2$) new dishes
- ...
- ith step: K dishes have been sampled, each by $n_1, \ldots, n_K$ customers; ith customer chooses jth dish *wp* n_j/i, plus Poisson(γ/i) new dishes.

Log growth for K_n:

$$K_n \sim \text{Poisson}(\gamma \log n).$$



[Image by M. Jordan]

Indian Buffet process Ghahramani and Griffiths [2006]

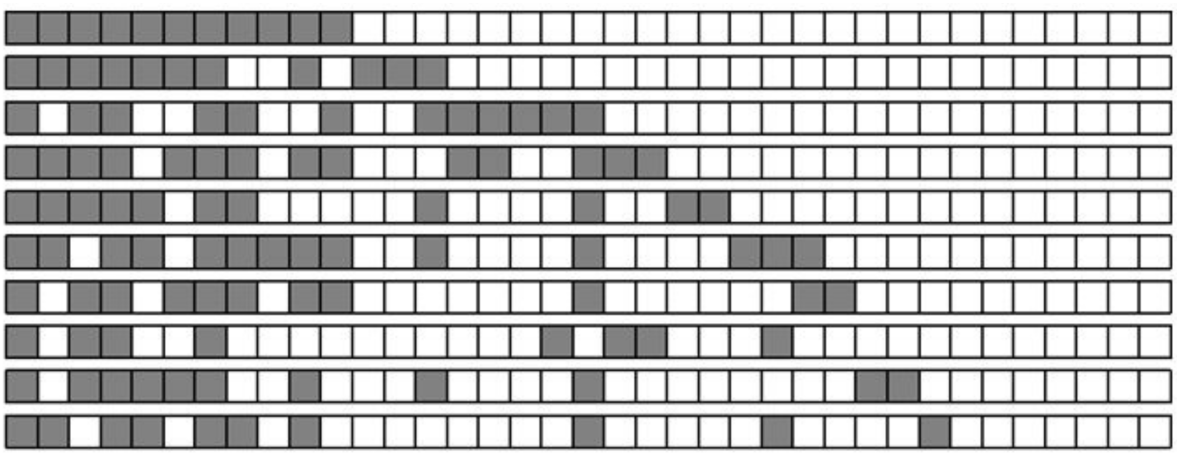
Feature allocation model: observations may share several features.

Generative model is as follows

- first customer samples Poisson(γ) dishes
- second customer chooses every dish of first customer *wp* $1/2$, plus Poisson($\gamma/2$) new dishes
- ...
- ith step: K dishes have been sampled, each by $n_1, \ldots, n_K$ customers; ith customer chooses jth dish *wp* n_j/i, plus Poisson(γ/i) new dishes.

Log growth for K_n:

$$K_n \sim \text{Poisson}(\gamma \log n).$$



[Image by M. Jordan]

Indian Buffet process Ghahramani and Griffiths [2006]

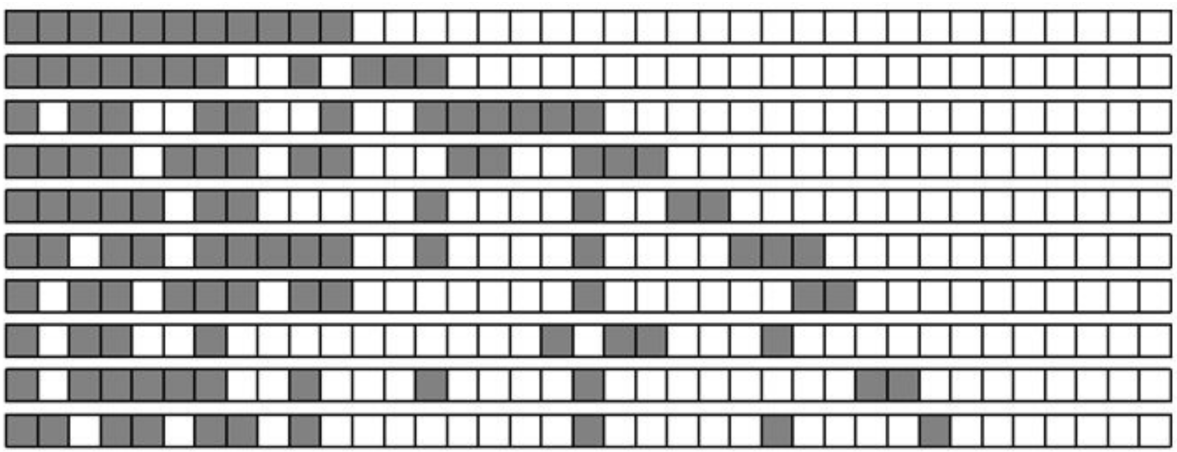
Feature allocation model: observations may share several features.

Generative model is as follows

- first customer samples Poisson(γ) dishes
- second customer chooses every dish of first customer *wp* $1/2$, plus Poisson($\gamma/2$) new dishes
- ...
- ith step: K dishes have been sampled, each by $n_1, \ldots, n_K$ customers; ith customer chooses jth dish *wp* n_j/i, plus Poisson(γ/i) new dishes.

Log growth for K_n:

$$K_n \sim \text{Poisson}(\gamma \log n).$$



[Image by M. Jordan]

Indian Buffet process Ghahramani and Griffiths [2006]

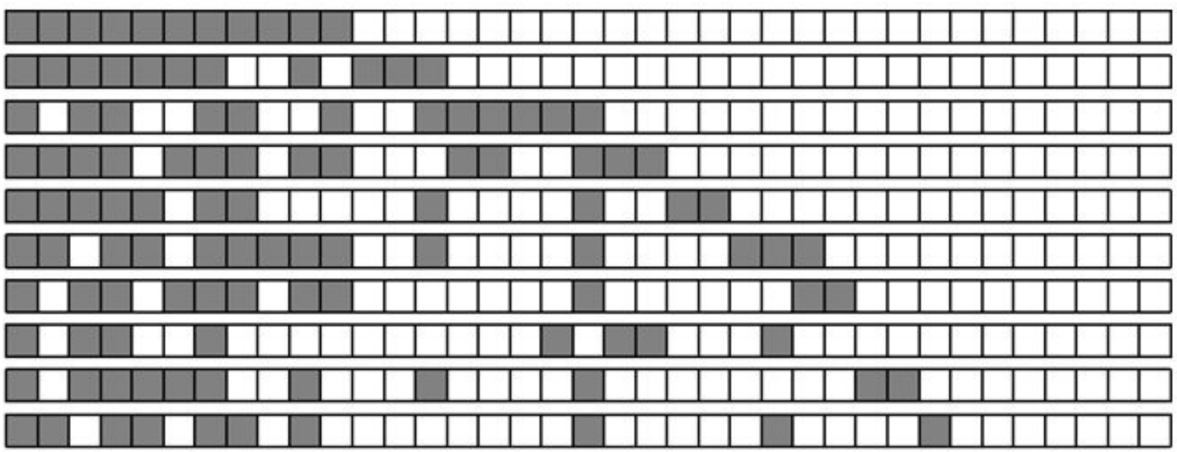
Feature allocation model: observations may share several features.

Generative model is as follows

- first customer samples Poisson(γ) dishes
- second customer chooses every dish of first customer *wp* $1/2$, plus Poisson($\gamma/2$) new dishes
- ...
- ith step: K dishes have been sampled, each by $n_1, \ldots, n_K$ customers; ith customer chooses jth dish *wp* n_j/i, plus Poisson(γ/i) new dishes.

Log growth for K_n:

$$K_n \sim \text{Poisson}(\gamma \log n).$$



[Image by M. Jordan]

Indian Buffet process Ghahramani and Griffiths [2006]

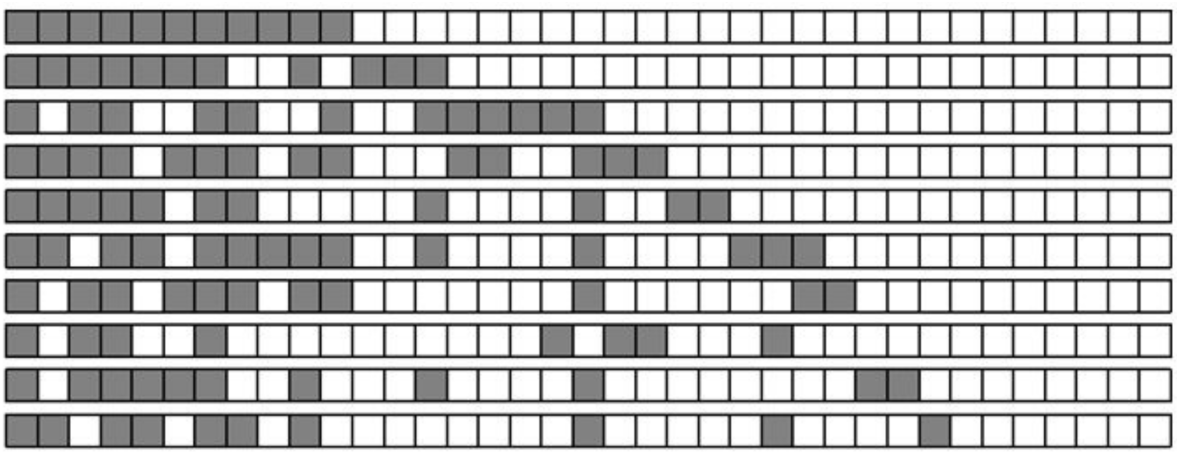
Feature allocation model: observations may share several features.

Generative model is as follows

- first customer samples Poisson(γ) dishes
- second customer chooses every dish of first customer *wp* $1/2$, plus Poisson($\gamma/2$) new dishes
- ...
- ith step: K dishes have been sampled, each by $n_1, \ldots, n_K$ customers; ith customer chooses jth dish *wp* n_j/i, plus Poisson(γ/i) new dishes.

Log growth for K_n:

$$K_n \sim \text{Poisson}(\gamma \log n).$$



[Image by M. Jordan]

Indian Buffet process Ghahramani and Griffiths [2006]

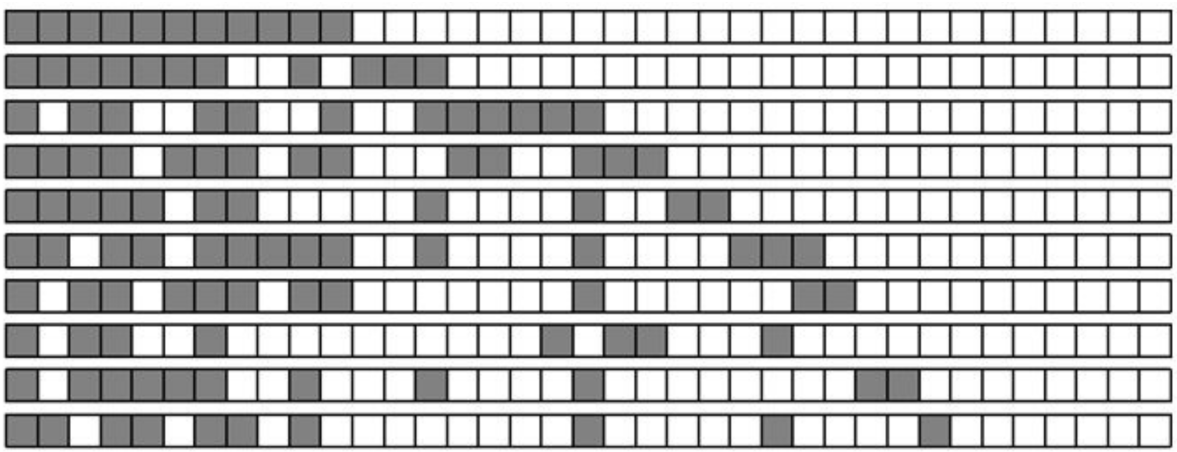
Feature allocation model: observations may share several features.

Generative model is as follows

- first customer samples Poisson(γ) dishes
- second customer chooses every dish of first customer *wp* $1/2$, plus Poisson($\gamma/2$) new dishes
- ...
- ith step: K dishes have been sampled, each by $n_1, \ldots, n_K$ customers; ith customer chooses jth dish *wp* n_j/i, plus Poisson(γ/i) new dishes.

Log growth for K_n:

$$K_n \sim \text{Poisson}(\gamma \log n).$$



[Image by M. Jordan]

Table of Contents

Discovery problem

- Population of individuals $(X_i)_{i\geq 1}$ belonging to an ideally infinite number of species $(\theta_j)_{j\geq 1}$, respective unknown proportions $(p_j)_{j\geq 1}$
- Given $X^n = (X_1, \ldots, X_n)$, make inference on the probability that X_{n+1} coincides with a species whose frequency is $\ell = 0, 1, \ldots, n$

$$\ell\text{-discovery} = \mathbb{P}(X_{n+1} \text{ is a species seen } \ell \text{ times in } X^n)$$

- Applications arising from ecology, biology, design of experiments, bioinformatics, genetics, linguistic, economics, network modeling, ...

Discovery problem

- Population of individuals $(X_i)_{i\geq 1}$ belonging to an ideally infinite number of species $(\theta_j)_{j\geq 1}$, respective unknown proportions $(p_j)_{j\geq 1}$
- Given $X^n = (X_1, \ldots, X_n)$, make inference on the probability that X_{n+1} coincides with a species whose frequency is $\ell = 0, 1, \ldots, n$

$$\ell\text{-discovery} = \mathbb{P}(X_{n+1} \text{ is a species seen } \ell \text{ times in } X^n)$$

- Applications arising from ecology, biology, design of experiments, bioinformatics, genetics, linguistic, economics, network modeling, ...

Discovery problem

- Population of individuals $(X_i)_{i\geq 1}$ belonging to an ideally infinite number of species $(\theta_j)_{j\geq 1}$, respective unknown proportions $(p_j)_{j\geq 1}$
- Given $X^n = (X_1, \ldots, X_n)$, make inference on the probability that X_{n+1} coincides with a species whose frequency is $\ell = 0, 1, \ldots, n$

$$\ell\text{-discovery} = \mathbb{P}(X_{n+1} \text{ is a species seen } \ell \text{ times in } X^n)$$

- Applications arising from ecology, biology, design of experiments, bioinformatics, genetics, linguistic, economics, network modeling, ...

Discovery problem, notations

- Given X^n, the ℓ-discovery probability is

$$D_n(\ell) = \sum_{j:\#(X_i=\theta_j)=\ell} p_j$$

- $D_n(0)$ denotes the proportion of yet unobserved species, or the probability of discovering a new species, or the missing mass
- Let $X_1^*, \ldots, X_{k_n}^*$ be the k_n distinct observations featured in X^n, with corresponding frequencies $(n_{1,n}, \ldots, n_{k_n,n})$
- The information provided by $(n_{1,n}, \ldots, n_{k_n,n})$ can be coded by $\mathbf{m}_n = (m_{1,n}, \ldots, m_{n,n})$ where $m_{\ell,n}$ = number of species in the sample $\boldsymbol{X}_n$ having frequency ℓ
  Under this alternative codification one obtains $\sum_{1\le\ell\le n} m_{\ell,n} = k_n$ and $\sum_{1\le\ell\le n} \ell m_{\ell,n} = n$.

Discovery problem, notations

- Given X^n, the ℓ-discovery probability is

$$D_n(\ell) = \sum_{j:\#(X_i=\theta_j)=\ell} p_j$$

- $D_n(0)$ denotes the proportion of yet unobserved species, or the probability of discovering a new species, or the missing mass
- Let $X_1^*, \ldots, X_{k_n}^*$ be the k_n distinct observations featured in X^n, with corresponding frequencies $(n_{1,n}, \ldots, n_{k_n,n})$
- The information provided by $(n_{1,n}, \ldots, n_{k_n,n})$ can be coded by $\mathbf{m}_n = (m_{1,n}, \ldots, m_{n,n})$ where $m_{\ell,n}$ = number of species in the sample $\boldsymbol{X}_n$ having frequency ℓ
  Under this alternative codification one obtains $\sum_{1\le\ell\le n} m_{\ell,n} = k_n$ and $\sum_{1\le\ell\le n} \ell m_{\ell,n} = n$.

Discovery problem, notations

- Given X^n, the ℓ-discovery probability is

$$D_n(\ell) = \sum_{j:\#(X_i=\theta_j)=\ell} p_j$$

- $D_n(0)$ denotes the proportion of yet unobserved species, or the probability of discovering a new species, or the missing mass
- Let $X_1^*, \ldots, X_{k_n}^*$ be the k_n distinct observations featured in X^n, with corresponding frequencies $(n_{1,n}, \ldots, n_{k_n,n})$
- The information provided by $(n_{1,n}, \ldots, n_{k_n,n})$ can be coded by $\mathbf{m}_n = (m_{1,n}, \ldots, m_{n,n})$ where $m_{\ell,n}$ = number of species in the sample $\boldsymbol{X}_n$ having frequency ℓ
  Under this alternative codification one obtains $\sum_{1\le\ell\le n} m_{\ell,n} = k_n$ and $\sum_{1\le\ell\le n} \ell m_{\ell,n} = n$.

Discovery problem, notations

- Given X^n, the ℓ-discovery probability is

$$D_n(\ell) = \sum_{j:\#(X_i=\theta_j)=\ell} p_j$$

- $D_n(0)$ denotes the proportion of yet unobserved species, or the probability of discovering a new species, or the missing mass
- Let $X_1^*, \ldots, X_{k_n}^*$ be the k_n distinct observations featured in X^n, with corresponding frequencies $(n_{1,n}, \ldots, n_{k_n,n})$
- The information provided by $(n_{1,n}, \ldots, n_{k_n,n})$ can be coded by $\mathbf{m}_n = (m_{1,n}, \ldots, m_{n,n})$ where $m_{\ell,n}$ = number of species in the sample $\boldsymbol{X}_n$ having frequency ℓ
  Under this alternative codification one obtains $\sum_{1\le\ell\le n} m_{\ell,n} = k_n$ and $\sum_{1\le\ell\le n} \ell m_{\ell,n} = n$.

Good–Turing estimators of discovery

Alan Turing and Irving John Good worked on this problem Bletchley Park to crack German ciphers for the Enigma machine during World War II. They proposed the empirical estimator

$$\check{\mathcal{D}}_n(0) = \frac{\text{Number of species observed once}}{\text{Total number of observations}}$$

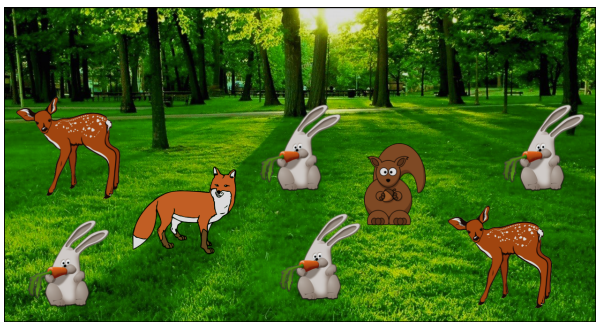
$$\check{\mathcal{D}}_n(0) = \frac{m_{1,n}}{n}, \quad \check{\mathcal{D}}_n(\ell) = \frac{(\ell+1)m_{\ell+1,n}}{n}, \quad \forall \ell \leq n$$

Good (1953)

Good–Turing estimators of discovery

Alan Turing and Irving John Good worked on this problem Bletchley Park to crack German ciphers for the Enigma machine during World War II. They proposed the empirical estimator

$$\check{\mathcal{D}}_n(0) = \frac{\text{Number of species observed once}}{\text{Total number of observations}}$$



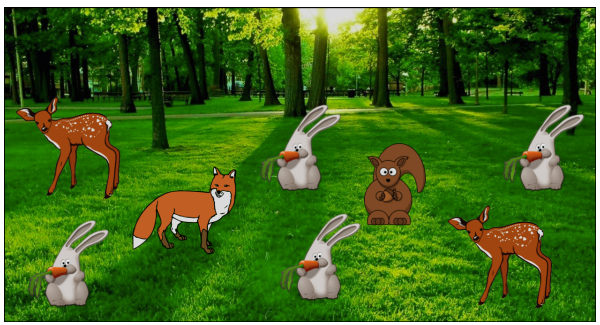
$$\check{\mathcal{D}}_n(0) = \frac{m_{1,n}}{n}, \quad \check{\mathcal{D}}_n(\ell) = \frac{(\ell+1)m_{\ell+1,n}}{n}, \quad \forall \ell \leq n$$

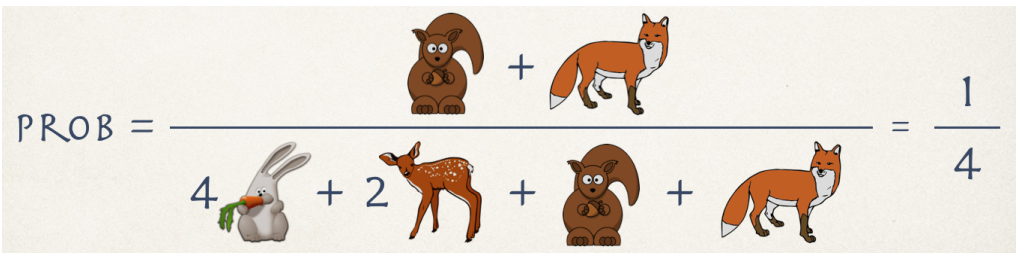
Good (1953)

Good–Turing estimators of discovery

Alan Turing and Irving John Good worked on this problem Bletchley Park to crack German ciphers for the Enigma machine during World War II. They proposed the empirical estimator

$$\check{\mathcal{D}}_n(0) = \frac{\text{Number of species observed once}}{\text{Total number of observations}}$$





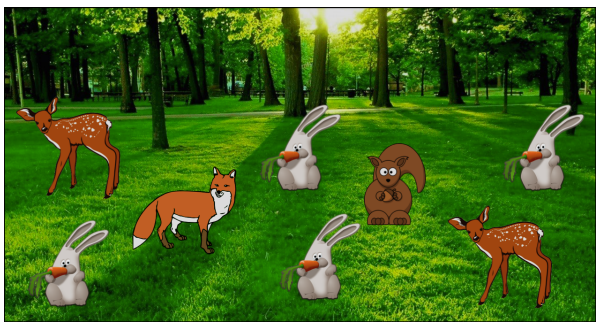
$$\check{\mathcal{D}}_n(0) = \frac{m_{1,n}}{n}, \quad \check{\mathcal{D}}_n(\ell) = \frac{(\ell+1)m_{\ell+1,n}}{n}, \quad \forall \ell \leq n$$

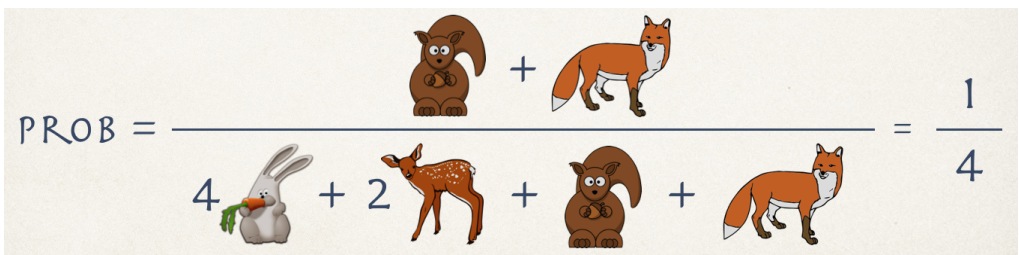
Good (1953)

Good–Turing estimators of discovery

Alan Turing and Irving John Good worked on this problem Bletchley Park to crack German ciphers for the Enigma machine during World War II. They proposed the empirical estimator

$$\check{\mathcal{D}}_n(0) = \frac{\text{Number of species observed once}}{\text{Total number of observations}}$$





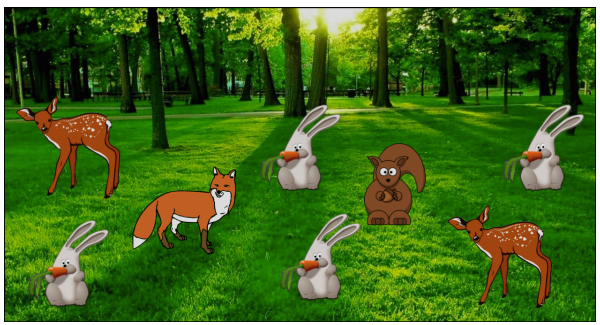
$$\check{\mathcal{D}}_n(0) = \frac{m_{1,n}}{n}, \quad \check{\mathcal{D}}_n(\ell) = \frac{(\ell+1)m_{\ell+1,n}}{n}, \quad \forall \ell \leq n$$

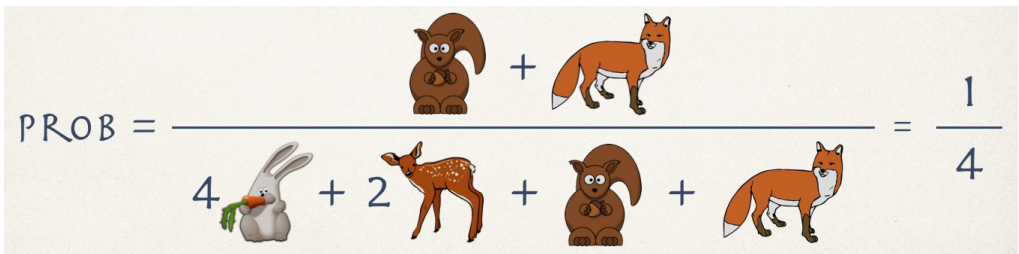
Good (1953)

Good–Turing estimators of discovery

Alan Turing and Irving John Good worked on this problem Bletchley Park to crack German ciphers for the Enigma machine during World War II. They proposed the empirical estimator

$$\check{\mathcal{D}}_n(0) = \frac{\text{Number of species observed once}}{\text{Total number of observations}}$$





$$\check{\mathcal{D}}_n(0) = \frac{m_{1,n}}{n}, \quad \check{\mathcal{D}}_n(\ell) = \frac{(\ell+1)m_{\ell+1,n}}{n}, \quad \forall \ell \leq n$$

Good (1953)

BNP counterparts of these empirical estimators with appropriate uncertainty quantification?

BNP model

BNP approach for estimating $D_n(\ell)$ based on randomization of unknown species proportions p_i's introduced by Lijoi, Mena and Prünster (2007)

Let $\mathscr{Q}$ denote a discrete random probability measure with random draws $P = \sum_{j\geq 1} p_j \delta_{\theta_j}$. Let $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ be a sample from a population with composition P, namely

$$\begin{aligned} X_i \mid P &\overset{\text{iid}}{\sim} P = \sum_{j\geq 1} p_j \delta_{\theta_j} \\ P &\sim \mathscr{Q} \end{aligned}$$

BNP model

BNP approach for estimating $D_n(\ell)$ based on randomization of unknown species proportions p_i's introduced by Lijoi, Mena and Prünster (2007)

Let $\mathscr{Q}$ denote a discrete random probability measure with random draws $P = \sum_{j\geq 1} p_j \delta_{\theta_j}$. Let $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ be a sample from a population with composition P, namely

$$\begin{aligned} X_i \mid P &\overset{\text{iid}}{\sim} P = \sum_{j\geq 1} p_j \delta_{\theta_j} \\ P &\sim \mathscr{Q} \end{aligned}$$

Define sets

$$\begin{aligned} A_0 &= \mathbb{X} \backslash \{X_1^*, \ldots, X_{k_n}^*\} \\ A_\ell &= \{X_j^* : n_{j,n} = \ell\} \end{aligned}$$

Unknown quantities reduce to

$$D_n(0) = P(A_0) \text{ and } D_n(\ell) = P(A_\ell)$$

BNP model

BNP approach for estimating $D_n(\ell)$ based on randomization of unknown species proportions p_i's introduced by Lijoi, Mena and Prünster (2007)

Let $\mathscr{Q}$ denote a discrete random probability measure with random draws $P = \sum_{j\geq 1} p_j \delta_{\theta_j}$. Let $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ be a sample from a population with composition P, namely

$$\begin{aligned} X_i \mid P &\overset{\text{iid}}{\sim} P = \sum_{j\geq 1} p_j \delta_{\theta_j} \\ P &\sim \mathscr{Q} \end{aligned}$$

Define sets

$$\begin{aligned} A_0 &= \mathbb{X} \backslash \{X_1^*, \ldots, X_{k_n}^*\} \\ A_\ell &= \{X_j^* : n_{j,n} = \ell\} \end{aligned}$$

Unknown quantities reduce to

$$D_n(0) = P(A_0) \text{ and } D_n(\ell) = P(A_\ell)$$

Which r.p.m. $\mathscr{Q}$ to choose?

BNP model

BNP approach for estimating $D_n(\ell)$ based on randomization of unknown species proportions p_i's introduced by Lijoi, Mena and Prünster (2007)

Let $\mathscr{Q}$ denote a discrete random probability measure with random draws $P = \sum_{j\geq 1} p_j \delta_{\theta_j}$. Let $\boldsymbol{X}_n = (X_1, \ldots, X_n)$ be a sample from a population with composition P, namely

$$X_i \mid P \overset{\text{iid}}{\sim} P = \sum_{j\geq 1} p_j \delta_{\theta_j}$$

$$P \sim \mathscr{Q}$$

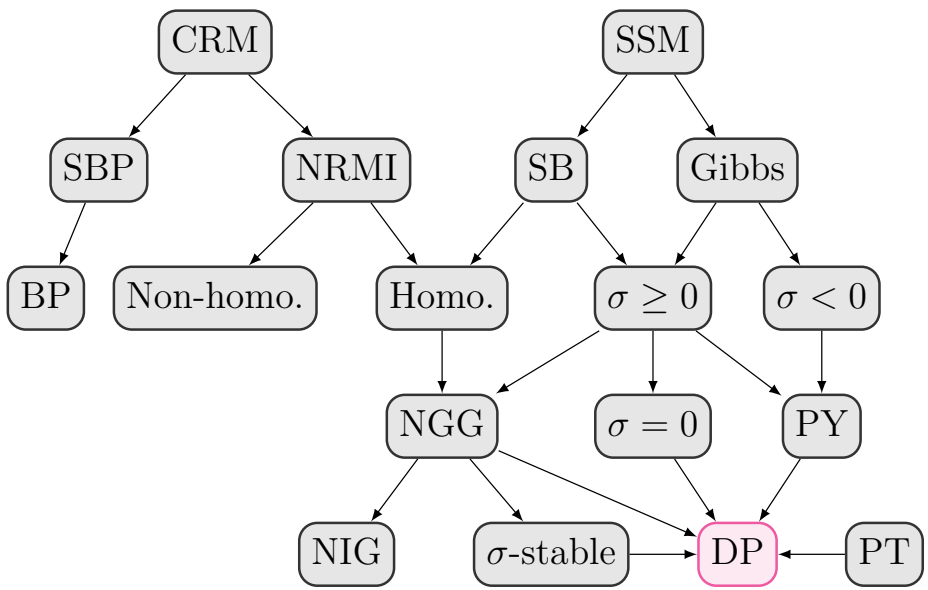
Define sets

$$A_0 = \mathbb{X} \backslash \{X_1^*, \ldots, X_{k_n}^*\}$$
$$A_\ell = \{X_j^* : n_{j,n} = \ell\}$$

Unknown quantities reduce to

$$D_n(0) = P(A_0) \text{ and } D_n(\ell) = P(A_\ell)$$

Which r.p.m. $\mathscr{Q}$ to choose?



BNP estimators of discovery

Re-write predictive distribution of Gibbs-type random probability measure

$$\mathbb{P}[X_{n+1} \in A \,|\, \boldsymbol{X}_n] = \frac{V_{n+1,k_n+1}}{V_{n,k_n}} G_0(A) + \frac{V_{n+1,k_n}}{V_{n,k_n}} \sum_{i=1}^{k_n} (n_{i,n} - \sigma)\, \delta_{X_i^*}(A)$$

BNP estimator of $D_n(\ell) = P(A_\ell)$ take the form of posterior expectations $\hat{\mathcal{D}}_n(\ell) = \mathbb{E}[P(A_\ell) \,|\, \boldsymbol{X}_n]$. Derived from the predictive using sets $A_0 = \mathbb{X} \backslash \{X_1^*, \ldots, X_{k_n}^*\}$ and $A_\ell = \{X_i^* : n_{i,n} = \ell\}$

BNP	Good–Turing
$\hat{\mathcal{D}}_n(0) = \frac{V_{n+1,k_n+1}}{V_{n,k_n}} \overset{\text{PY}}{=} \frac{\alpha+\sigma k_n}{\alpha+n}$	$\check{\mathcal{D}}_n(0) = \frac{m_{1,n}}{n}$
$\hat{\mathcal{D}}_n(\ell) = (\ell-\sigma) m_{\ell,n} \frac{V_{n+1,k_n}}{V_{n,k_n}} \overset{\text{PY}}{=} (\ell-\sigma)\frac{m_{\ell,n}}{\alpha+n}$	$\check{\mathcal{D}}_n(\ell) = (\ell+1)\frac{m_{\ell+1,n}}{n}$

The most notable difference between $\hat{\mathcal{D}}_n(\ell)$ and $\check{\mathcal{D}}_n(\ell)$ consists in the use of the information contained in the sample $\boldsymbol{X}_n$: $\check{\mathcal{D}}_n(\ell)$ is a function of $m_{\ell+1,n}$, and not on $(k_n, m_{\ell,n})$ as one would intuitively expect for an estimator of the ℓ-discovery

BNP estimators of discovery

Re-write predictive distribution of Gibbs-type random probability measure

$$\mathbb{P}[X_{n+1} \in A \,|\, \boldsymbol{X}_n] = \frac{V_{n+1,k_n+1}}{V_{n,k_n}} G_0(A) + \frac{V_{n+1,k_n}}{V_{n,k_n}} \sum_{i=1}^{k_n} (n_{i,n} - \sigma)\, \delta_{X_i^*}(A)$$

BNP estimator of $D_n(\ell) = P(A_\ell)$ take the form of posterior expectations $\hat{\mathcal{D}}_n(\ell) = \mathbb{E}[P(A_\ell) \,|\, \boldsymbol{X}_n]$. Derived from the predictive using sets $A_0 = \mathbb{X} \backslash \{X_1^*, \ldots, X_{k_n}^*\}$ and $A_\ell = \{X_i^* : n_{i,n} = \ell\}$

BNP	Good–Turing
$\hat{\mathcal{D}}_n(0) = \frac{V_{n+1,k_n+1}}{V_{n,k_n}} \overset{\text{PY}}{=} \frac{\alpha+\sigma k_n}{\alpha+n}$	$\check{\mathcal{D}}_n(0) = \frac{m_{1,n}}{n}$
$\hat{\mathcal{D}}_n(\ell) = (\ell-\sigma) m_{\ell,n} \frac{V_{n+1,k_n}}{V_{n,k_n}} \overset{\text{PY}}{=} (\ell-\sigma)\frac{m_{\ell,n}}{\alpha+n}$	$\check{\mathcal{D}}_n(\ell) = (\ell+1)\frac{m_{\ell+1,n}}{n}$

The most notable difference between $\hat{\mathcal{D}}_n(\ell)$ and $\check{\mathcal{D}}_n(\ell)$ consists in the use of the information contained in the sample $\boldsymbol{X}_n$: $\check{\mathcal{D}}_n(\ell)$ is a function of $m_{\ell+1,n}$, and not on $(k_n, m_{\ell,n})$ as one would intuitively expect for an estimator of the ℓ-discovery

BNP estimators of discovery

Re-write predictive distribution of Gibbs-type random probability measure

$$\mathbb{P}[X_{n+1} \in A \,|\, \boldsymbol{X}_n] = \frac{V_{n+1,k_n+1}}{V_{n,k_n}} G_0(A) + \frac{V_{n+1,k_n}}{V_{n,k_n}} \sum_{i=1}^{k_n} (n_{i,n} - \sigma)\, \delta_{X_i^*}(A)$$

BNP estimator of $D_n(\ell) = P(A_\ell)$ take the form of posterior expectations $\hat{\mathcal{D}}_n(\ell) = \mathbb{E}[P(A_\ell) \,|\, \boldsymbol{X}_n]$. Derived from the predictive using sets $A_0 = \mathbb{X} \backslash \{X_1^*, \ldots, X_{k_n}^*\}$ and $A_\ell = \{X_i^* : n_{i,n} = \ell\}$

BNP	Good–Turing
$\hat{\mathcal{D}}_n(0) = \frac{V_{n+1,k_n+1}}{V_{n,k_n}} \overset{\text{PY}}{=} \frac{\alpha+\sigma k_n}{\alpha+n}$	$\check{\mathcal{D}}_n(0) = \frac{m_{1,n}}{n}$
$\hat{\mathcal{D}}_n(\ell) = (\ell-\sigma) m_{\ell,n} \frac{V_{n+1,k_n}}{V_{n,k_n}} \overset{\text{PY}}{=} (\ell-\sigma)\frac{m_{\ell,n}}{\alpha+n}$	$\check{\mathcal{D}}_n(\ell) = (\ell+1)\frac{m_{\ell+1,n}}{n}$

The most notable difference between $\hat{\mathcal{D}}_n(\ell)$ and $\check{\mathcal{D}}_n(\ell)$ consists in the use of the information contained in the sample $\boldsymbol{X}_n$: $\check{\mathcal{D}}_n(\ell)$ is a function of $m_{\ell+1,n}$, and not on $(k_n, m_{\ell,n})$ as one would intuitively expect for an estimator of the ℓ-discovery

BNP estimators of discovery

Re-write predictive distribution of Gibbs-type random probability measure

$$\mathbb{P}[X_{n+1} \in A \,|\, \boldsymbol{X}_n] = \frac{V_{n+1,k_n+1}}{V_{n,k_n}} G_0(A) + \frac{V_{n+1,k_n}}{V_{n,k_n}} \sum_{i=1}^{k_n} (n_{i,n} - \sigma)\, \delta_{X_i^*}(A)$$

BNP estimator of $D_n(\ell) = P(A_\ell)$ take the form of posterior expectations $\hat{\mathcal{D}}_n(\ell) = \mathbb{E}[P(A_\ell) \,|\, \boldsymbol{X}_n]$. Derived from the predictive using sets $A_0 = \mathbb{X} \backslash \{X_1^*, \ldots, X_{k_n}^*\}$ and $A_\ell = \{X_i^* : n_{i,n} = \ell\}$

BNP	Good–Turing
$\hat{\mathcal{D}}_n(0) = \frac{V_{n+1,k_n+1}}{V_{n,k_n}} \overset{\text{PY}}{=} \frac{\alpha+\sigma k_n}{\alpha+n}$	$\check{\mathcal{D}}_n(0) = \frac{m_{1,n}}{n}$
$\hat{\mathcal{D}}_n(\ell) = (\ell-\sigma) m_{\ell,n} \frac{V_{n+1,k_n}}{V_{n,k_n}} \overset{\text{PY}}{=} (\ell-\sigma)\frac{m_{\ell,n}}{\alpha+n}$	$\check{\mathcal{D}}_n(\ell) = (\ell+1)\frac{m_{\ell+1,n}}{n}$

The most notable difference between $\hat{\mathcal{D}}_n(\ell)$ and $\check{\mathcal{D}}_n(\ell)$ consists in the use of the information contained in the sample $\boldsymbol{X}_n$: $\check{\mathcal{D}}_n(\ell)$ is a function of $m_{\ell+1,n}$, and not on $(k_n, m_{\ell,n})$ as one would intuitively expect for an estimator of the ℓ-discovery

Credible intervals for discovery

- Arbel et al. [2017]; Arbel and Favaro [2018] Under specification of Pitman–Yor process prior, the posterior distribution of $D_n(\ell)$ is a simple Beta distribution

$$D_n(0) = P(A_0) \,|\, \boldsymbol{X}_n \sim B_{\alpha+\sigma k_n, n-\sigma k_n}$$

and

$$D_n(\ell) = P(A_\ell) \,|\, \boldsymbol{X}_n \sim B_{(\ell-\sigma)m_{\ell,n}, \alpha+n-(\ell-\sigma)m_{\ell,n}}$$

- Similar results in the general Gibbs class
- Practical tool for deriving credible intervals for the BNP estimator $\hat{\mathcal{D}}_n(\ell)$, for any $\ell = 0, 1, \ldots, n$: numerical evaluation of appropriate quantiles of the distribution of $P(A_\ell) \,|\, \boldsymbol{X}_n$

Credible intervals for discovery

- Arbel et al. [2017]; Arbel and Favaro [2018] Under specification of Pitman–Yor process prior, the posterior distribution of $D_n(\ell)$ is a simple Beta distribution

$$D_n(0) = P(A_0) \,|\, \boldsymbol{X}_n \sim B_{\alpha+\sigma k_n, n-\sigma k_n}$$

and

$$D_n(\ell) = P(A_\ell) \,|\, \boldsymbol{X}_n \sim B_{(\ell-\sigma)m_{\ell,n}, \alpha+n-(\ell-\sigma)m_{\ell,n}}$$

- Similar results in the general Gibbs class
- Practical tool for deriving credible intervals for the BNP estimator $\hat{\mathcal{D}}_n(\ell)$, for any $\ell = 0, 1, \ldots, n$: numerical evaluation of appropriate quantiles of the distribution of $P(A_\ell) \,|\, \boldsymbol{X}_n$

Credible intervals for discovery

- Arbel et al. [2017]; Arbel and Favaro [2018] Under specification of Pitman–Yor process prior, the posterior distribution of $D_n(\ell)$ is a simple Beta distribution

$$D_n(0) = P(A_0) \,|\, \boldsymbol{X}_n \sim B_{\alpha+\sigma k_n, n-\sigma k_n}$$

and

$$D_n(\ell) = P(A_\ell) \,|\, \boldsymbol{X}_n \sim B_{(\ell-\sigma)m_{\ell,n}, \alpha+n-(\ell-\sigma)m_{\ell,n}}$$

- Similar results in the general Gibbs class
- Practical tool for deriving credible intervals for the BNP estimator $\hat{\mathcal{D}}_n(\ell)$, for any $\ell = 0, 1, \ldots, n$: numerical evaluation of appropriate quantiles of the distribution of $P(A_\ell) \,|\, \boldsymbol{X}_n$

Credible intervals: sketch of proof

Steps:

1. Compute by induction the posterior moments of $P(A)$ for any set A
2. Evaluate the expressions on the sets A_0 and A_ℓ to get posterior moments of $D_n(0)$ and $D_n(\ell)$

$$\mathbb{E}[(P(A_0))^r \mid \boldsymbol{X}_n] = \sum_{i=0}^{r} \binom{r}{i} (-1)^i \frac{V_{n+i,k_n}}{V_{n,k_n}} (n - \sigma k_n)_i$$

$$\mathbb{E}[(P(A_\ell))^r \mid \boldsymbol{X}_n] = \frac{V_{n+r,k_n}}{V_{n,k_n}} ((\ell - \sigma) m_{\ell,n})_r$$

3. Identify the moments with moments of known distributions
   - Beta for Pitman–Yor
   - Mixture of simple r.v. otherwise

Credible intervals: sketch of proof

Steps:

1. Compute by induction the posterior moments of $P(A)$ for any set A
2. Evaluate the expressions on the sets A_0 and A_ℓ to get posterior moments of $D_n(0)$ and $D_n(\ell)$

$$\mathbb{E}[(P(A_0))^r \mid \boldsymbol{X}_n] = \sum_{i=0}^{r} \binom{r}{i} (-1)^i \frac{V_{n+i,k_n}}{V_{n,k_n}} (n - \sigma k_n)_i$$

$$\mathbb{E}[(P(A_\ell))^r \mid \boldsymbol{X}_n] = \frac{V_{n+r,k_n}}{V_{n,k_n}} ((\ell - \sigma) m_{\ell,n})_r$$

3. Identify the moments with moments of known distributions
   - Beta for Pitman–Yor
   - Mixture of simple r.v. otherwise

Credible intervals: sketch of proof

Steps:

1. Compute by induction the posterior moments of $P(A)$ for any set A
2. Evaluate the expressions on the sets A_0 and A_ℓ to get posterior moments of $D_n(0)$ and $D_n(\ell)$

$$\mathbb{E}[(P(A_0))^r \mid \boldsymbol{X}_n] = \sum_{i=0}^{r} \binom{r}{i} (-1)^i \frac{V_{n+i,k_n}}{V_{n,k_n}} (n - \sigma k_n)_i$$

$$\mathbb{E}[(P(A_\ell))^r \mid \boldsymbol{X}_n] = \frac{V_{n+r,k_n}}{V_{n,k_n}} ((\ell - \sigma) m_{\ell,n})_r$$

3. Identify the moments with moments of known distributions
   - Beta for Pitman–Yor
   - Mixture of simple r.v. otherwise

Credible intervals: sketch of proof

Steps:

1. Compute by induction the posterior moments of $P(A)$ for any set A
2. Evaluate the expressions on the sets A_0 and A_ℓ to get posterior moments of $D_n(0)$ and $D_n(\ell)$

$$\mathbb{E}[(P(A_0))^r \mid \boldsymbol{X}_n] = \sum_{i=0}^{r} \binom{r}{i} (-1)^i \frac{V_{n+i,k_n}}{V_{n,k_n}} (n - \sigma k_n)_i$$

$$\mathbb{E}[(P(A_\ell))^r \mid \boldsymbol{X}_n] = \frac{V_{n+r,k_n}}{V_{n,k_n}} ((\ell - \sigma) m_{\ell,n})_r$$

3. Identify the moments with moments of known distributions
   - Beta for Pitman–Yor
   - Mixture of simple r.v. otherwise

Credible intervals: sketch of proof

Steps:

1. Compute by induction the posterior moments of $P(A)$ for any set A
2. Evaluate the expressions on the sets A_0 and A_ℓ to get posterior moments of $D_n(0)$ and $D_n(\ell)$

$$\mathbb{E}[(P(A_0))^r \mid \boldsymbol{X}_n] = \sum_{i=0}^{r} \binom{r}{i} (-1)^i \frac{V_{n+i,k_n}}{V_{n,k_n}} (n - \sigma k_n)_i$$

$$\mathbb{E}[(P(A_\ell))^r \mid \boldsymbol{X}_n] = \frac{V_{n+r,k_n}}{V_{n,k_n}} ((\ell - \sigma) m_{\ell,n})_r$$

3. Identify the moments with moments of known distributions
   - Beta for Pitman–Yor
   - Mixture of simple r.v. otherwise

Application to EST libraries

Application to genomic datasets called Expressed Sequence Tags (EST) libraries for unicellular organisms sequenced recently

- *Naegleria gruberi* aerobic library consists of $n = 959$ ESTs with $k_n = 473$ distinct genes and $m_{\ell,959} = 346, 57, 19, 12, 9, 5, 4, 2, 4, 5, 4, 1, 1, 1, 1, 1, 1,$ for $\ell \in \{1, 2, \ldots, 12\} \cup \{16, 17, 18\} \cup \{27\} \cup \{55\}$
- *Naegleria gruberi* anaerobic library consists of $n = 969$ ESTs with $k_n = 631$ distinct genes and $m_{\ell,969} = 491, 72, 30, 9, 13, 5, 3, 1, 2, 0, 1, 0, 1,$ for $\ell \in \{1, 2, \ldots, 13\}$

Prior specification: Pitman–Yor process, with empirical Bayes procedure for estimating (σ, α)

- $\hat{\sigma} = 0.67$, $\hat{\alpha} = 46$ for the *Naegleria gruberi* aerobic library
- $\hat{\sigma} = 0.65$, $\hat{\alpha} = 155$ for the *Naegleria gruberi* anaerobic library

Application to EST libraries

Application to genomic datasets called Expressed Sequence Tags (EST) libraries for unicellular organisms sequenced recently

- *Naegleria gruberi* aerobic library consists of $n = 959$ ESTs with $k_n = 473$ distinct genes and $m_{\ell,959} = 346, 57, 19, 12, 9, 5, 4, 2, 4, 5, 4, 1, 1, 1, 1, 1, 1,$ for $\ell \in \{1, 2, \ldots, 12\} \cup \{16, 17, 18\} \cup \{27\} \cup \{55\}$
- *Naegleria gruberi* anaerobic library consists of $n = 969$ ESTs with $k_n = 631$ distinct genes and $m_{\ell,969} = 491, 72, 30, 9, 13, 5, 3, 1, 2, 0, 1, 0, 1,$ for $\ell \in \{1, 2, \ldots, 13\}$

Prior specification: Pitman–Yor process, with empirical Bayes procedure for estimating (σ, α)

- $\hat{\sigma} = 0.67$, $\hat{\alpha} = 46$ for the *Naegleria gruberi* aerobic library
- $\hat{\sigma} = 0.65$, $\hat{\alpha} = 155$ for the *Naegleria gruberi* anaerobic library

Application to EST libraries

Application to genomic datasets called Expressed Sequence Tags (EST) libraries for unicellular organisms sequenced recently

- *Naegleria gruberi* aerobic library consists of $n = 959$ ESTs with $k_n = 473$ distinct genes and $m_{\ell,959} = 346, 57, 19, 12, 9, 5, 4, 2, 4, 5, 4, 1, 1, 1, 1, 1, 1,$ for $\ell \in \{1, 2, \ldots, 12\} \cup \{16, 17, 18\} \cup \{27\} \cup \{55\}$
- *Naegleria gruberi* anaerobic library consists of $n = 969$ ESTs with $k_n = 631$ distinct genes and $m_{\ell,969} = 491, 72, 30, 9, 13, 5, 3, 1, 2, 0, 1, 0, 1,$ for $\ell \in \{1, 2, \ldots, 13\}$

Prior specification: Pitman–Yor process, with empirical Bayes procedure for estimating (σ, α)

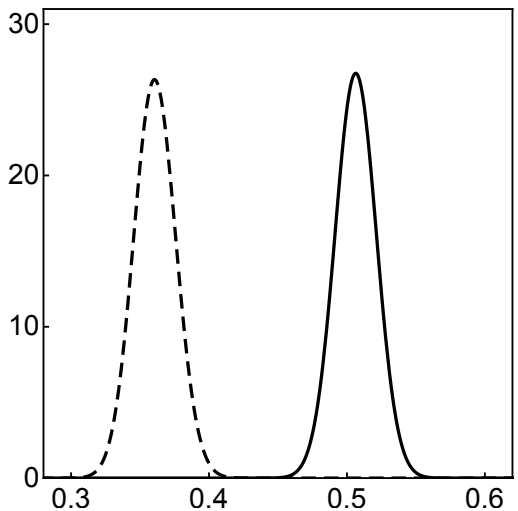
- $\hat{\sigma} = 0.67$, $\hat{\alpha} = 46$ for the *Naegleria gruberi* aerobic library
- $\hat{\sigma} = 0.65$, $\hat{\alpha} = 155$ for the *Naegleria gruberi* anaerobic library

Application to EST libraries

Posterior distributions of discovery probabilities $D_n(\ell)$, for $\ell \in \{0, 1, 5\}$: dashed curve for aerobic, solid curve for anaerobic

Application to EST libraries

Posterior distributions of discovery probabilities $D_n(\ell)$, for $\ell \in \{0, 1, 5\}$: dashed curve for aerobic, solid curve for anaerobic



Application to EST libraries

Posterior distributions of discovery probabilities $D_n(\ell)$, for $\ell \in \{0, 1, 5\}$: dashed curve for aerobic, solid curve for anaerobic

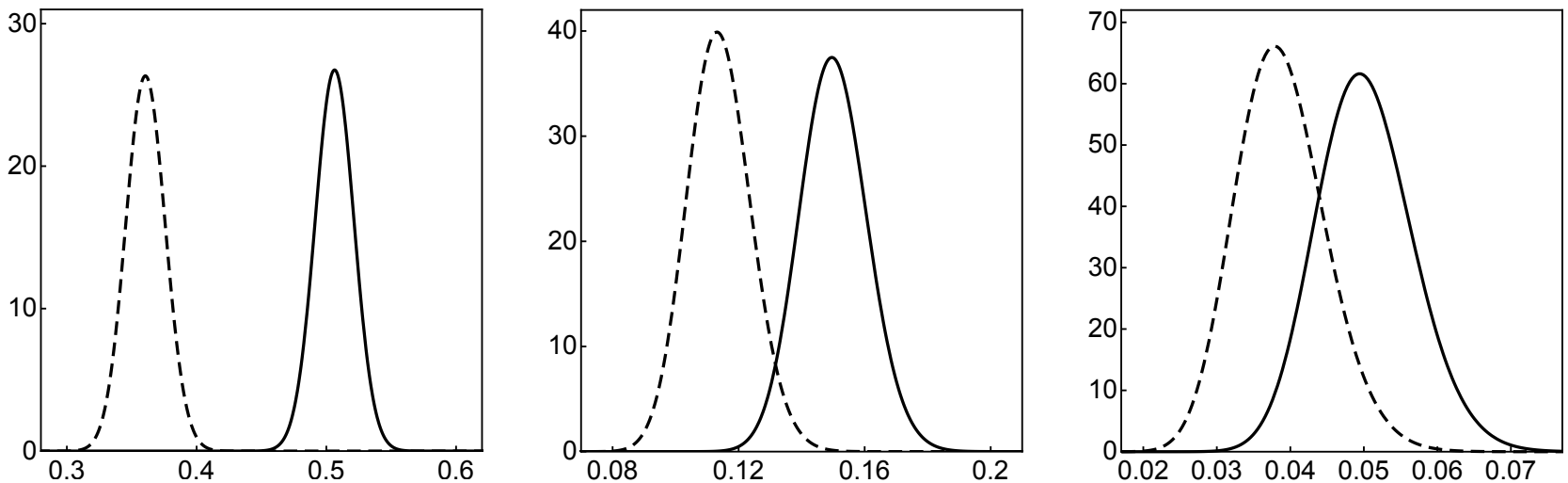


Table of Contents

Some research directions in BNP

Methodological

- Mixture models: density, regression, survival analysis, clustering
- Feature allocation models: Indian Buffet process
- Hierarchical model: hierarchical Dirichlet process
- Dependent priors (dependent Dirichlet process), multivariate processes, non exchangeable data, copulas
- Gaussian processes
- Model selection, high dimensional setting, sparsity
- Relational data, networks, random graphs
- Structured data, segmentation of images, graphs, time series
- Extreme value theory
- ...

Some research directions in BNP

Methodological

- Mixture models: density, regression, survival analysis, clustering
- Feature allocation models: Indian Buffet process
- Hierarchical model: hierarchical Dirichlet process
- Dependent priors (dependent Dirichlet process), multivariate processes, non exchangeable data, copulas
- Gaussian processes
- Model selection, high dimensional setting, sparsity
- Relational data, networks, random graphs
- Structured data, segmentation of images, graphs, time series
- Extreme value theory
- ...

Some research directions in BNP

Methodological

- Mixture models: density, regression, survival analysis, clustering
- Feature allocation models: Indian Buffet process
- Hierarchical model: hierarchical Dirichlet process
- Dependent priors (dependent Dirichlet process), multivariate processes, non exchangeable data, copulas
- Gaussian processes
- Model selection, high dimensional setting, sparsity
- Relational data, networks, random graphs
- Structured data, segmentation of images, graphs, time series
- Extreme value theory
- ...

Some research directions in BNP

Methodological

- Mixture models: density, regression, survival analysis, clustering
- Feature allocation models: Indian Buffet process
- Hierarchical model: hierarchical Dirichlet process
- Dependent priors (dependent Dirichlet process), multivariate processes, non exchangeable data, copulas
- Gaussian processes
- Model selection, high dimensional setting, sparsity
- Relational data, networks, random graphs
- Structured data, segmentation of images, graphs, time series
- Extreme value theory
- ...

Some research directions in BNP

Methodological

- Mixture models: density, regression, survival analysis, clustering
- Feature allocation models: Indian Buffet process
- Hierarchical model: hierarchical Dirichlet process
- Dependent priors (dependent Dirichlet process), multivariate processes, non exchangeable data, copulas
- Gaussian processes
- Model selection, high dimensional setting, sparsity
- Relational data, networks, random graphs
- Structured data, segmentation of images, graphs, time series
- Extreme value theory
- ...

Some research directions in BNP

Methodological

- Mixture models: density, regression, survival analysis, clustering
- Feature allocation models: Indian Buffet process
- Hierarchical model: hierarchical Dirichlet process
- Dependent priors (dependent Dirichlet process), multivariate processes, non exchangeable data, copulas
- Gaussian processes
- Model selection, high dimensional setting, sparsity
- Relational data, networks, random graphs
- Structured data, segmentation of images, graphs, time series
- Extreme value theory
- ...

Some research directions in BNP

Methodological

- Mixture models: density, regression, survival analysis, clustering
- Feature allocation models: Indian Buffet process
- Hierarchical model: hierarchical Dirichlet process
- Dependent priors (dependent Dirichlet process), multivariate processes, non exchangeable data, copulas
- Gaussian processes
- Model selection, high dimensional setting, sparsity
- Relational data, networks, random graphs
- Structured data, segmentation of images, graphs, time series
- Extreme value theory
- ...

Some research directions in BNP

Methodological

- Mixture models: density, regression, survival analysis, clustering
- Feature allocation models: Indian Buffet process
- Hierarchical model: hierarchical Dirichlet process
- Dependent priors (dependent Dirichlet process), multivariate processes, non exchangeable data, copulas
- Gaussian processes
- Model selection, high dimensional setting, sparsity
- Relational data, networks, random graphs
- Structured data, segmentation of images, graphs, time series
- Extreme value theory
- ...

Some research directions in BNP

Methodological

- Mixture models: density, regression, survival analysis, clustering
- Feature allocation models: Indian Buffet process
- Hierarchical model: hierarchical Dirichlet process
- Dependent priors (dependent Dirichlet process), multivariate processes, non exchangeable data, copulas
- Gaussian processes
- Model selection, high dimensional setting, sparsity
- Relational data, networks, random graphs
- Structured data, segmentation of images, graphs, time series
- Extreme value theory
- ...

Some research directions in BNP

Methodological

- Mixture models: density, regression, survival analysis, clustering
- Feature allocation models: Indian Buffet process
- Hierarchical model: hierarchical Dirichlet process
- Dependent priors (dependent Dirichlet process), multivariate processes, non exchangeable data, copulas
- Gaussian processes
- Model selection, high dimensional setting, sparsity
- Relational data, networks, random graphs
- Structured data, segmentation of images, graphs, time series
- Extreme value theory
- ...

Some research directions in BNP

Theoretical

- Theoretical validation, eg asymptotic behavior of the posterior: consistency, rates of convergence, Bernstein–von Mises theorem
- Support properties for priors: full support
- Model misspecification
- ...

Some research directions in BNP

Theoretical

- Theoretical validation, eg asymptotic behavior of the posterior: consistency, rates of convergence, Bernstein–von Mises theorem
- Support properties for priors: full support
- Model misspecification
- ...

Some research directions in BNP

Theoretical

- Theoretical validation, eg asymptotic behavior of the posterior: consistency, rates of convergence, Bernstein–von Mises theorem
- Support properties for priors: full support
- Model misspecification
- . . .

Some research directions in BNP

Theoretical

- Theoretical validation, eg asymptotic behavior of the posterior: consistency, rates of convergence, Bernstein–von Mises theorem
- Support properties for priors: full support
- Model misspecification
- ...

Some research directions in BNP

Computational

- Markov chain Monte Carlo algorithms
- Scalable algorithms in the context of Big Data: variational inference
- Links with Machine Learning, Bayesian Deep Learning and Deep Bayesian Learning
- ...

Some research directions in BNP

Computational

- Markov chain Monte Carlo algorithms
- Scalable algorithms in the context of Big Data: variational inference
- Links with Machine Learning, Bayesian Deep Learning and Deep Bayesian Learning
- ...

Some research directions in BNP

Computational

- Markov chain Monte Carlo algorithms
- Scalable algorithms in the context of Big Data: variational inference
- Links with Machine Learning, Bayesian Deep Learning and Deep Bayesian Learning
- . . .

Some research directions in BNP

Computational

- Markov chain Monte Carlo algorithms
- Scalable algorithms in the context of Big Data: variational inference
- Links with Machine Learning, Bayesian Deep Learning and Deep Bayesian Learning
- ...

Some research directions in BNP

Applications

- Biostatistics
- Environmental science, ecotoxicology
- Neurosciences, neuroimaging
- Astrophysics
- Linguistics
- ...

Some research directions in BNP

Applications

- Biostatistics
- Environmental science, ecotoxicology
- Neurosciences, neuroimaging
- Astrophysics
- Linguistics
- ...

Some research directions in BNP

Applications

- Biostatistics
- Environmental science, ecotoxicology
- Neurosciences, neuroimaging
- Astrophysics
- Linguistics
- ...

Some research directions in BNP

Applications

- Biostatistics
- Environmental science, ecotoxicology
- Neurosciences, neuroimaging
- Astrophysics
- Linguistics
- ...

Some research directions in BNP

Applications

- Biostatistics
- Environmental science, ecotoxicology
- Neurosciences, neuroimaging
- Astrophysics
- Linguistics
- . . .

Some research directions in BNP

Applications

- Biostatistics
- Environmental science, ecotoxicology
- Neurosciences, neuroimaging
- Astrophysics
- Linguistics
- ...

References I

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, pages 1152–1174.

Arbel, J., Corradin, R., and Nipoti, B. (2018a). Dirichlet process mixtures under affine transformations of the data. *Submitted*.

Arbel, J., De Blasi, P., and Prünster, I. (2018b). Stochastic approximations to the Pitman–Yor process. *Bayesian Analysis*.

Arbel, J. and Favaro, S. (2018). Approximating predictive probabilities of Gibbs-type priors. *Submitted*.

Arbel, J., Favaro, S., Nipoti, B., and Teh, Y. W. (2017). Bayesian nonparametric inference for discovery probabilities: credible intervals and large sample asymptotics. *Statistica Sinica*, 27:839–858.

Arbel, J., Gayraud, G., and Rousseau, J. (2013). Bayesian optimal adaptive estimation using a sieve prior. *Scandinavian Journal of Statistics*, 40(3):549–570.

Arbel, J., Lijoi, A., and Nipoti, B. (2016a). Full Bayesian inference with hazard mixture models. *Computational Statistics & Data Analysis*, 93:359–372.

References II

Arbel, J., Mengersen, K., Raymond, B., Winsley, T., and King, C. (2015). Application of a Bayesian nonparametric model to derive toxicity estimates based on the response of Antarctic microbial communities to fuel contaminated soil. *Ecology and Evolution*, 5(13):2633–2645.

Arbel, J., Mengersen, K., and Rousseau, J. (2016b). Bayesian nonparametric dependent model for partially replicated data: the influence of fuel spills on species diversity. *Annals of Applied Statistics*, 10(3):1496–1516.

Arbel, J. and Prünster, I. (2017). A moment-matching Ferguson & Klass algorithm. *Statistics and Computing*, 27(1):3–17.

Barrios, E., Lijoi, A., Nieto-Barajas, L. E., and Prünster, I. (2013). Modeling with normalized random measure mixture models. *Statistical Science*, 28(3):313–334.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Brix, A. (1999). Generalized gamma measures and shot-noise Cox processes. *Advances in Applied Probability*, pages 929–953.

Clauset, A., Shalizi, C. R., and Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review*, 51(4):661–703.

References III

Dahl, D. B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. *Bayesian inference for gene expression and proteomics*, pages 201–218.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(2):212–229.

Ewens, W. J. (1972). The sampling theory of selectively neutral alleles. *Theoretical population biology*, 3(1):87–112.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230.

Ferguson, T. S. and Klass, M. J. (1972). A representation of independent increment processes without gaussian components. *The Annals of Mathematical Statistics*, 43(5):1634–1643.

Ghahramani, Z. and Griffiths, T. L. (2006). Infinite latent feature models and the Indian buffet process. In *Advances in neural information processing systems*, pages 475–482.

Ishwaran, H. and James, L. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173.

References IV

Jara, A., Hanson, T., Quintana, F., Müller, P., and Rosner, G. (2011). DPpackage: Bayesian non-and semi-parametric modelling in R. *Journal of statistical software*, 40(5):1.

Lawless, C. and Arbel, J. (2018). A simple proof of Pitman–Yor's Chinese restaurant process from its stick-breaking representation. *Preprint*.

Lü, H., Arbel, J., and Forbes, F. (2018). Bayesian Nonparametric Priors for Hidden Markov Random Fields. *Preprint*.

Marchal, O. and Arbel, J. (2017). On the sub-Gaussianity of the Beta and Dirichlet distributions. *Electronic Communications in Probability*, 22:1–14.

Meilă, M. (2007). Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895.

Miller, J. W. and Harrison, M. T. (2013). A simple example of Dirichlet process mixture inconsistency for the number of components. In *Advances in neural information processing systems*, pages 199–206.

Muliere, P. and Tardella, L. (1998). Approximating distributions of random functionals of ferguson-dirichlet priors. *Canadian Journal of Statistics*, 26(2):283–297.

References V

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265.

Newman, M. E. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary physics*, 46(5):323–351.

Papaspiliopoulos, O. and Roberts, G. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, 95(1):169.

Rajkowski, Ł. (2016). Analysis of MAP in CRP Normal-Normal model. *arXiv preprint arXiv:1606.03275*.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.

Teh, Y., Jordan, M., Beal, M., and Blei, D. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Wade, S. and Ghahramani, Z. (2018). Bayesian cluster analysis: Point estimation and credible balls. *Bayesian Analysis*.

Walker, S. G. (2007). Sampling the dirichlet mixture model with slices. *Communications in Statistics—Simulation and Computation®*, 36(1):45–54.