Bayesian nonparametric estimation of the probability of discovering new species

Lijoi, Mena & Prünster (2007)

Stefano Cortinovis

Monday 31st January, 2022

Species sampling framework

We formalize the process of drawing samples from a large population of individuals that can be grouped in different species as follows:

Species sampling framework

We formalize the process of drawing samples from a large population of individuals that can be grouped in different species as follows:

- Let $(X_n)_{n \geq 1}$ be a sequence of random variables taking values in some set $\mathbb{X}$.
  - X_n represents the **species** of the n-th individual sampled
  - $\mathbb{X}$ represents an arbitrary set of **tags** used to label species

Species sampling framework

We formalize the process of drawing samples from a large population of individuals that can be grouped in different species as follows:

- Let $(X_n)_{n \geq 1}$ be a sequence of random variables taking values in some set $\mathbb{X}$.
  - X_n represents the **species** of the n-th individual sampled
  - $\mathbb{X}$ represents an arbitrary set of **tags** used to label species
- Define

$$M_j := \begin{cases} 1 & \text{if } j = 1 \\ \inf\{n \colon n > M_{j-1}, X_n \notin \{X_1, \ldots, X_{n-1}\}\} & \text{if } j \geq 2 \end{cases}$$

and, for $M_j < \infty$, let $\tilde{X}_j := X_{M_j}$.

- $\tilde{X}_j$ represents the j-th **distinct species** to be observed

Species sampling framework

Formalize the process of drawing samples from a large population of individuals of various species as follows:

- Let $K_n := \max\{j : k \leq n \text{ and } M_j < \infty\}$.
  - K_n represents the **number of distinct species** to appear in the first n observations

Species sampling framework

Formalize the process of drawing samples from a large population of individuals of various species as follows:

- Let $K_n := \max\{j : k \leq n \text{ and } M_j < \infty\}$.
  - K_n represents the **number of distinct species** to appear in the first n observations
- Define

$$N_{j,n} := \sum_{i=1}^{n} \mathbb{1}(X_i = \tilde{X}_j)$$

for $j = 1, ..., K_n$ and let $\mathrm{N}_n = (N_{1,n}, ..., N_{K_n,n})$.

  - $N_{j,n}$ represents the **number of times** that the j-th species $\tilde{X}_j$ appears in the first n observations

Species sampling problem

Taking inspiration from biological and ecological studies, given a sample of size n containing j distinct species, denoted by $X_j^{(1,n)}$, we are interested in determining:

Species sampling problem

Taking inspiration from biological and ecological studies, given a sample of size n containing j distinct species, denoted by $X_j^{(1,n)}$, we are interested in determining:

1. The probability distribution of the **number of new species** recorded among the following m observations
   - Denote the second unobserved sample of size m by $X^{(2,m)} = (X_{n+1}, ..., X_{n+m})$
   - Denote the number of new species in $X^{(2,m)}$ by $K_m^{(n)} = K_{n+m} - K_n$
   - Then, 1 amounts to determining $\text{pr}(K_m^{(n)} = k | X_j^{(1,n)})$ for $k = 0, ..., m$ and for $j = 1, ..., n$

Species sampling problem

Taking inspiration from biological and ecological studies, given a sample of size n containing j distinct species, denoted by $X_j^{(1,n)}$, we are interested in determining:

1. The probability distribution of the **number of new species** recorded among the following m observations
   - Denote the second unobserved sample of size m by $X^{(2,m)} = (X_{n+1}, ..., X_{n+m})$
   - Denote the number of new species in $X^{(2,m)}$ by $K_m^{(n)} = K_{n+m} - K_n$
   - Then, 1 amounts to determining $\text{pr}(K_m^{(n)} = k | X_j^{(1,n)})$ for $k = 0, ..., m$ and for $j = 1, ..., n$
2. The probability of observing a **new species** at the $(n + m + 1)$-th draw
   - Using the notation introduced above, 2 amounts to determining the **random** probability $D_m^{n:j} := \text{pr}(K_1^{(m+n)} = 1 | X_j^{(1,n)}, X^{(2,m)})$

Species Sampling Model (Pitman, 1996)

We call $(X_n)_{n\geq 1}$ a sample from random distribution $\tilde{P}$ if $X_1, X_2, ...|\tilde{P} \overset{iid}{\sim} \tilde{P}$.

Definition 1 (Species sampling model)

Let $(X_n)_{n\geq 1}$ be a sample from a **discrete** random distribution $\tilde{P}$ of the form

$$\tilde{P} = \sum_{i=1}^{\infty} P_i \delta_{\hat{X}_i} \tag{1}$$

where $(P_i)_{i\geq 1}$ is a sequence of random variables such that $P_i \geq 0$ a.s. for every i, $\sum_i P_i = 1$ a.s., and $(\hat{X}_i)_{n\geq 1} \overset{iid}{\sim} \nu$ independently of (P_i) for ν diffuse.

The setup above, with $\tilde{P}$ as in (1) and $(X_n)_{n\geq 1}|\tilde{P} \overset{iid}{\sim} \tilde{P}$ is called a (proper) *species sampling model* (**SSM**) with *species sampling process* (**SSP**) $\tilde{P}$.

SSM Characterization

As a result of the discrete nature of $\tilde{P}$, an SSP induces an **infinite exchangeable random partition** (Pitman, 1995).

SSM Characterization

As a result of the discrete nature of $\tilde{P}$, an SSP induces an **infinite exchangeable random partition** (Pitman, 1995).

The **strong link** between a species sampling model and the partition it induces was unveiled by Pitman, who showed that the former is **characterized** by ν and by the exchangeable partition probability function (**EPPF**) associated with the latter.

In other words, for a given ν, an SSM is characterized by the joint distributions of K_n and N_n, i.e.

$$\mathrm{pr}[\{K_n = k\} \cap \{N_{j,n} = n_j,\ j = 1, ..., k\}],$$

for $n \geq 1$.

Gibbs-type priors (Gnedin & Pitman, 2006)

An important class of SSPs, particularly attractive for the **simplicity** of their EPPF, is represented by Gibbs-type priors.

Definition 2 (Gibbs-type prior)

A Gibbs-type prior of order $\sigma \in (0,1)$ is a SSP that induces an infinite exchangeable random partition with EPPF of the form

$$\text{pr}[\{K_n = k\} \cap \{N_{jn} = n_j,\ j = 1, ..., k\}] = V_{n,k} \prod_{j=1}^{k} (1-\sigma)_{n_j - 1}$$

for some set of non-negative weights $\{V_{n,k} : n \geq 1, 1 \leq k \leq n\}$ satisfying the recurrence relation

$$V_{n,k} = (n - \sigma) V_{n+1,k} + V_{n+1,k+1}. \tag{2}$$

Predictive distribution

A Gibbs-type prior is also characterized by its **predictive distributions**, which take the form

1. $P(X_1 \in \cdot) = \nu(\cdot)$
2. $P(X_{n+1} \in \cdot | X^{(n)}) = \frac{V_{n+1,k+1}}{V_{n,k}} \nu(\cdot) + \frac{V_{n+1,k}}{V_{n,k}} \sum_{j=1}^{k} (n_j - \sigma) \delta_{\tilde{X}_j}(A)$

where $X_1, ..., X_n$ is a sample of size n containing $K_n = k$ distinct species with frequencies $n_1, ..., n_k$.

Predictive distribution

A Gibbs-type prior is also characterized by its **predictive distributions**, which take the form

1. $P(X_1 \in \cdot) = \nu(\cdot)$
2. $P(X_{n+1} \in \cdot | X^{(n)}) = \frac{V_{n+1,k+1}}{V_{n,k}} \nu(\cdot) + \frac{V_{n+1,k}}{V_{n,k}} \sum_{j=1}^{k} (n_j - \sigma) \delta_{\tilde{X}_j}(A)$

where $X_1, ..., X_n$ is a sample of size n containing $K_n = k$ distinct species with frequencies $n_1, ..., n_k$.

The expression above sheds some light on the **inferential implications** connected with a Gibbs-type prior. Indeed, given that $X^{(n)}$ has been observed, the probability of observing a **new** species **only** depends on n and k.

Moreover, given that a **new** species is observed, it is assigned a **label** according to ν. On the other hand, given that an **old** species is observed, the probability that its label corresponds to $\tilde{X}_j$ depends on n_j and σ.

Prior distribution

Proposition 1 (Gnedin & Pitman, 2006)

Let $(X_n)_{n\geq 1}$ be a sequence of exchangeable observations governed by a Gibbs-type prior of order σ and with a set of non-negative weights $\{V_{n,k}: n \geq 1, 1 \leq k \leq n\}$ satisfying (2). Then, for any $k \in \{1, ..., n\}$,

$$pr(K_n = k) = \frac{V_{n,k}}{\sigma^k} \mathscr{C}(n, k; \sigma),$$

where $\mathscr{C}(n, k; \sigma)$ is a generalized factorial coefficient.

Posterior distribution

Proposition 2 (Lijoi, Mena & Prünster, 2007)

Let $(X_n)_{n\geq 1}$ be a sequence of exchangeable observations governed by a Gibbs-type prior of order σ and with a set of non-negative weights $\{V_{n,k} : n \geq 1, 1 \leq k \leq n\}$ satisfying (2). Then, for any $k \in \{0, ..., m\}$ and for any $j \in \{1, ..., n\}$,

$$pr(K_m^{(n)} = k | X_j^{(1,n)}) = \frac{V_{n+m,j+k}}{V_{n,j}} \frac{1}{\sigma^k} \mathscr{C}(m, k; \sigma, -n + j\sigma),$$

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

Notice that, in order to compute the probability above, the only information that we need about the sample $X_j^{(1,n)}$ other than its size is the number K_n of distinct species observed. We refer to this structural feature of Gibbs-type priors as **sufficiency** of K_n.

Discovery probability at the $(n+m+1)$-th draw

Thanks to the sufficiency of K_n, we have that

$$D_m^{(n:j)} = \text{pr}(K_1^{(m+n)} = 1 | X_j^{(1,n)}, X^{(2,m)}) = \text{pr}(K_1^{(m+n)} = 1 | K_n, K_m^{(n)}).$$

Proposition 3 (Lijoi, Mena & Prünster, 2007)

Let $(X_n)_{n\geq 1}$ be a sequence of exchangeable observations governed by a Gibbs-type prior of order σ and with a set of non-negative weights $\{V_{n,k} : n \geq 1, 1 \leq k \leq n\}$ satisfying (2). Then, the Bayes estimate under a squared loss function of $D_m^{(n:j)}$ is

$$\hat{D}_m^{(n:j)} = \sum_{k=0}^{m} \frac{V_{n+m+1,j+k+1}}{V_{n,j}} \frac{1}{\sigma^k} \mathscr{C}(m, k; \sigma, -n + j\sigma),$$

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

Dirichlet process (Ferguson, 1973)

The Dirichlet process $\tilde{P} \sim \text{DP}(\alpha)$ is a Gibbs-type prior for $\sigma \to 0$ and with

$$V_{n,k} = \frac{\theta^k}{(\theta)_n}$$

where $\theta = \alpha(\mathbb{X})$. Then, using the results presented before, we have

$$\text{pr}(K_n = k) = \frac{\theta^k}{(\theta)_n} |s(n,k)|$$

$$\text{pr}(K_m^{(n)} = k | X_j^{(1,n)}) = \frac{\theta^k (\theta)_n}{(\theta)_{n+m}} \sum_{l=k}^{m} \binom{m}{l} |s(l,k)| (n)_{m-l}$$

$$\hat{D}_m^{(n:j)} = \frac{\theta}{(\theta+n)_{m+1}} \sum_{k=0}^{m} \theta^k \sum_{l=k}^{m} \binom{m}{l} |s(l,k)| (n)_{m-l}$$

where $|s(n,k)|$ is an unsigned Stirling number of the first kind.

Two-parameter Poisson-Dirichlet process (Pitman, 1995)

The Poisson-Dirichlet process $\tilde{P} \sim \text{PD}(\sigma, \theta)$ is a Gibbs-type prior of order σ with

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

Then, using the results presented before, we have

$$\text{pr}(K_n = k) = \frac{\prod_{i=1}^{k-1}(\theta + i\sigma)}{\sigma^k (\theta+1)_{n-1}} \mathscr{C}(n, k; \sigma)$$

$$\text{pr}(K_m^{(n)} = k | X_j^{(1,n)}) = \frac{(\theta+1)_{n-1}}{(\theta+1)_{n+m-1}} \frac{\prod_{i=j}^{j+k-1}(\theta + i\sigma)}{\sigma^k} \mathscr{C}(m, k; \sigma, -n + j\sigma)$$

$$\hat{D}_m^{(n:j)} = \frac{(\theta+1)_{n-1}}{(\theta+1)_{n+m}} \sum_{k=0}^{m} \frac{\prod_{i=j}^{j+k}(\theta + i\sigma)}{\sigma^k} \mathscr{C}(m, k; \sigma, -n + j\sigma)$$

Normalized inverse Gaussian process (Lijoi et al., 2005)

The normalized inverse Gaussian process $\tilde{P} \sim \text{N-IG}(\theta)$ is a Gibbs-type prior of order $\sigma = 1/2$ with

$$V_{n,k} = \frac{e^{\theta}(-\theta^2)^{n-1}}{\Gamma(n)2^{k-1}} \sum_{i=0}^{n-1} \binom{n-1}{i} (-\theta^2)^{-i} \Gamma(k+2+2i-2n;\theta)$$

where $\Gamma(\nu, x)$ is the incomplete gamma function.

Once again, we can apply the results presented before to obtain the desired quantities. For the sake of brevity, we don't show them here, but, as in the other cases, they are available in **closed form** and their calculation requires **little computational effort**.

Synthetic example - Prior distribution

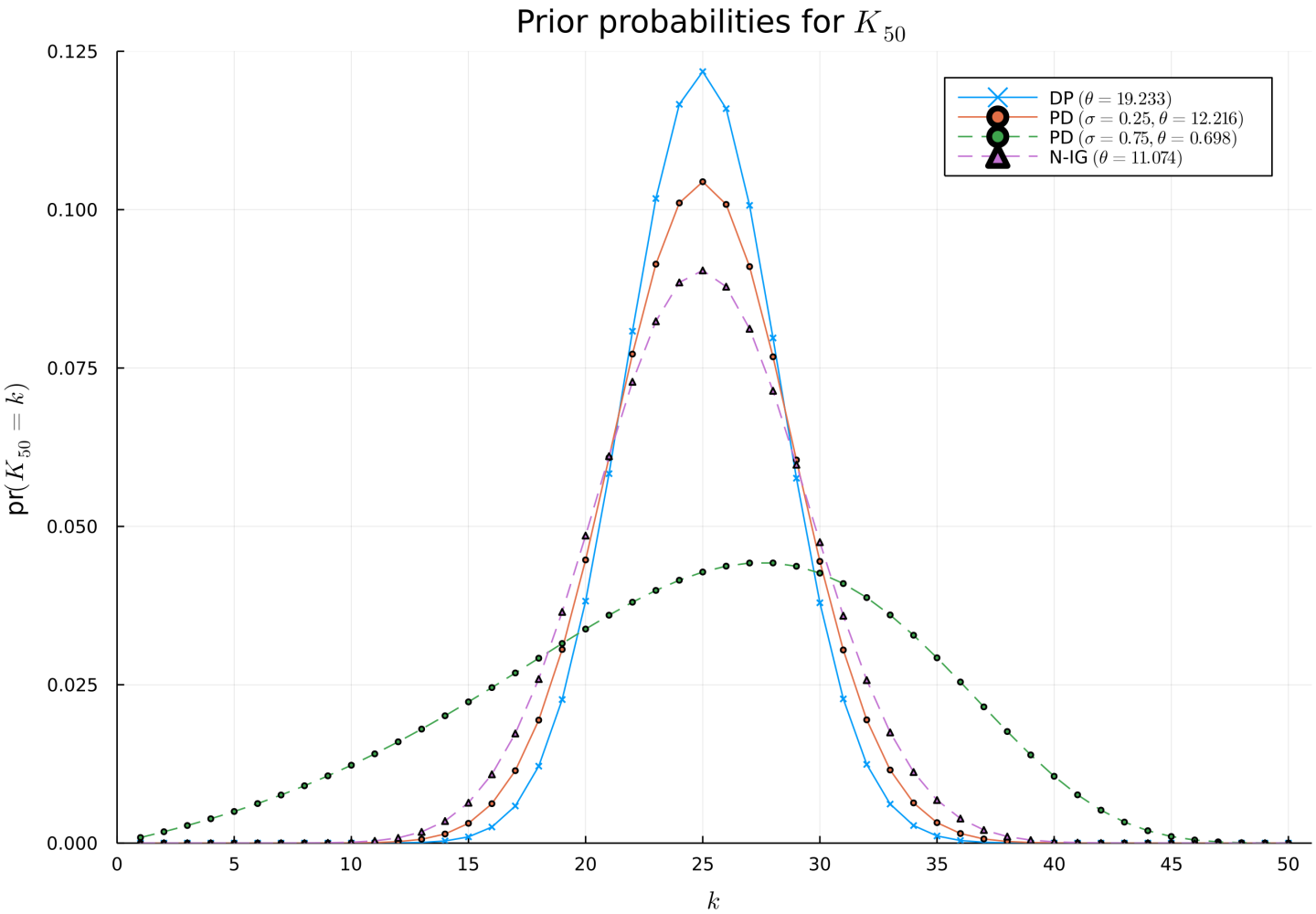

Figure: Prior probabilities of K_{50} corresponding to four different choices of Gibbs-type priors such that $E(K_{50}) = 25$.

Synthetic example - Posterior distribution

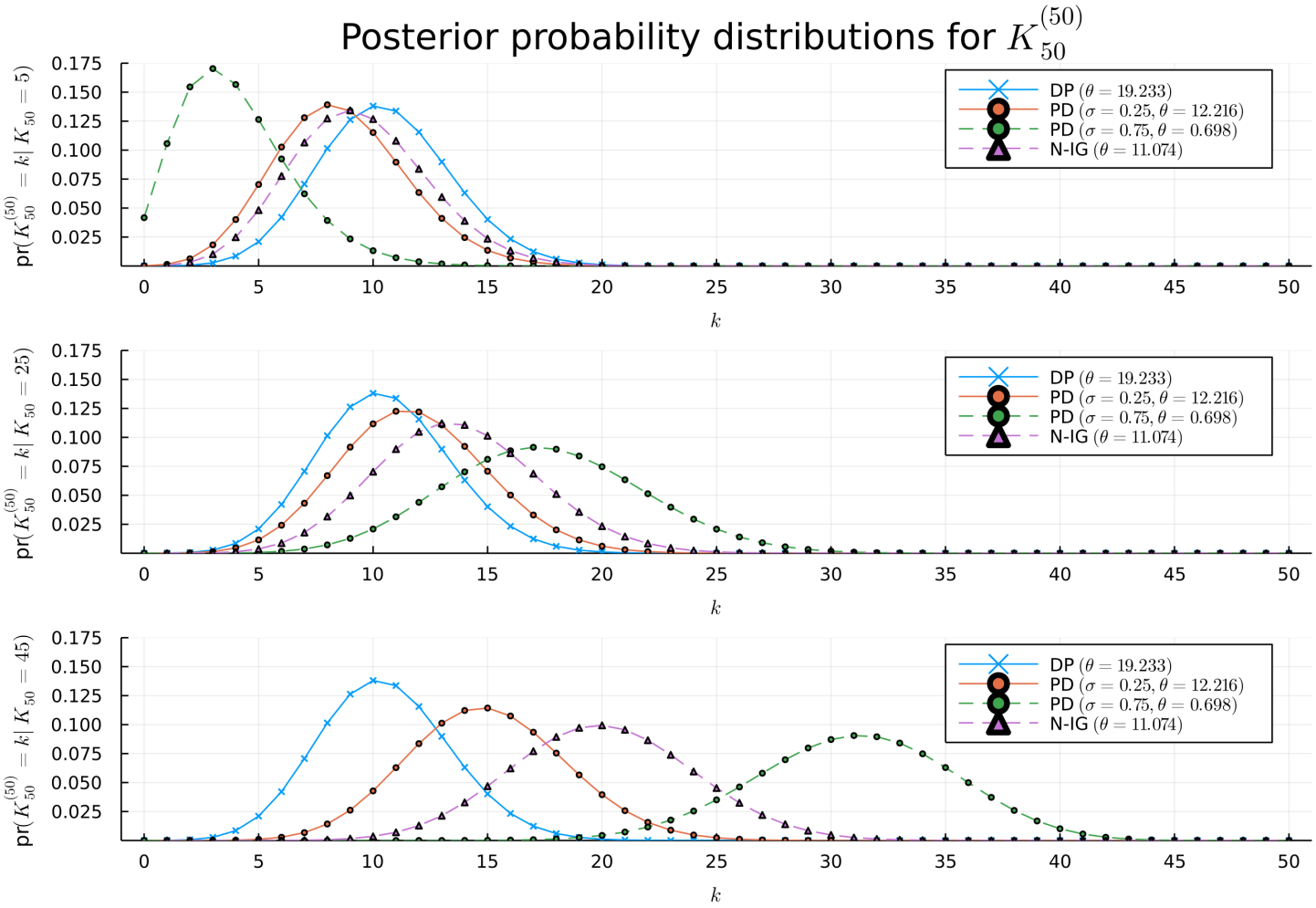

Figure: Posterior probabilities of $(K_{50}^{(50)} | K_{50} = j)$ for $j \in \{5, 25, 45\}$ corresponding to four different choices of Gibbs-type priors.

Genomics example

We consider a sample from a cDNA library made from buds of tomato flowers (Mao, 2004):

- Sample of size $n = 2586$
- Number of distinct genes $j = 1825$
- Number of clusters by size

$$r_i = 1434, 253, 71, 33, 11, 6, 2, 3, 1, 2, 2, 1, 1, 1, 2, 1, 1$$

for $i \in \{1, ..., 14\} \cup \{16, 23, 27\}$

and two choices of PD(σ, θ) process as priors:

1. Maximum likelihood choice of $(\sigma, \theta) = (0.612, 741)$, i.e.

$$(0.612, 741) = \underset{\sigma,\theta}{\text{argmax}}\ \phi_{n_1,...,n_j}(\sigma, \theta)$$

   where $\phi_{n_1,...,n_j}(\sigma, \theta)$ is the EPPF of the prior evaluated at $(j, n_1, ..., n_j)$ seen as a function of σ and θ

2. Common choice of $\sigma = 0.5$ combined with θ such that $E(K_{2586}) = 1825$, i.e. $\theta = 1093.313$

Genomics example - $\hat{D}_m^{2586:1825}$ and comparisons

Table 3. *Genomics example. Estimates in percentages for* $m \in \{517, 1034, 1552, 2069, 2586\}$ *obtained with the estimator* $\hat{D}_m^{(2586:1825)}$ *arising from the two choices of the Poisson–Dirichlet process, from the moment–based estimator* $\hat{U}_e$ *and from the likelihood–based estimator* $\tilde{U}_e$*. For the Poisson–Dirichlet processes, the 95% highest posterior density intervals are also shown.*

m	Poisson–Dirichlet process $(\sigma, \theta) = (0{\cdot}612, 741)$	Poisson–Dirichlet process $(\sigma, \theta) = (0{\cdot}5, 1093{\cdot}313)$	$\hat{U}_e$	$\tilde{U}_e$
0	55·84	54·52	55·46	55·45
517	52·80 ∈ (52·42, 53·19)	51·04 ∈ (50·76, 51·33)	51·86	51·83
1034	50·28 ∈ (49·79, 50·77)	48·17 ∈ (47·80, 48·53)	48·74	48·66
1552	48·14 ∈ (47·59, 48·69)	45·72 ∈ (45·31, 46·13)	45·99	45·74
2069	46·30 ∈ (45·70, 46·88)	43·62 ∈ (43·18, 44·05)	43·51	42·80
2586	44·68 ∈ (44·06, 45·30)	41·78 ∈ (41·32, 42·23)	41·24	39·98

Genomics example - $\hat{D}_m^{2586:1825}$ and comparisons

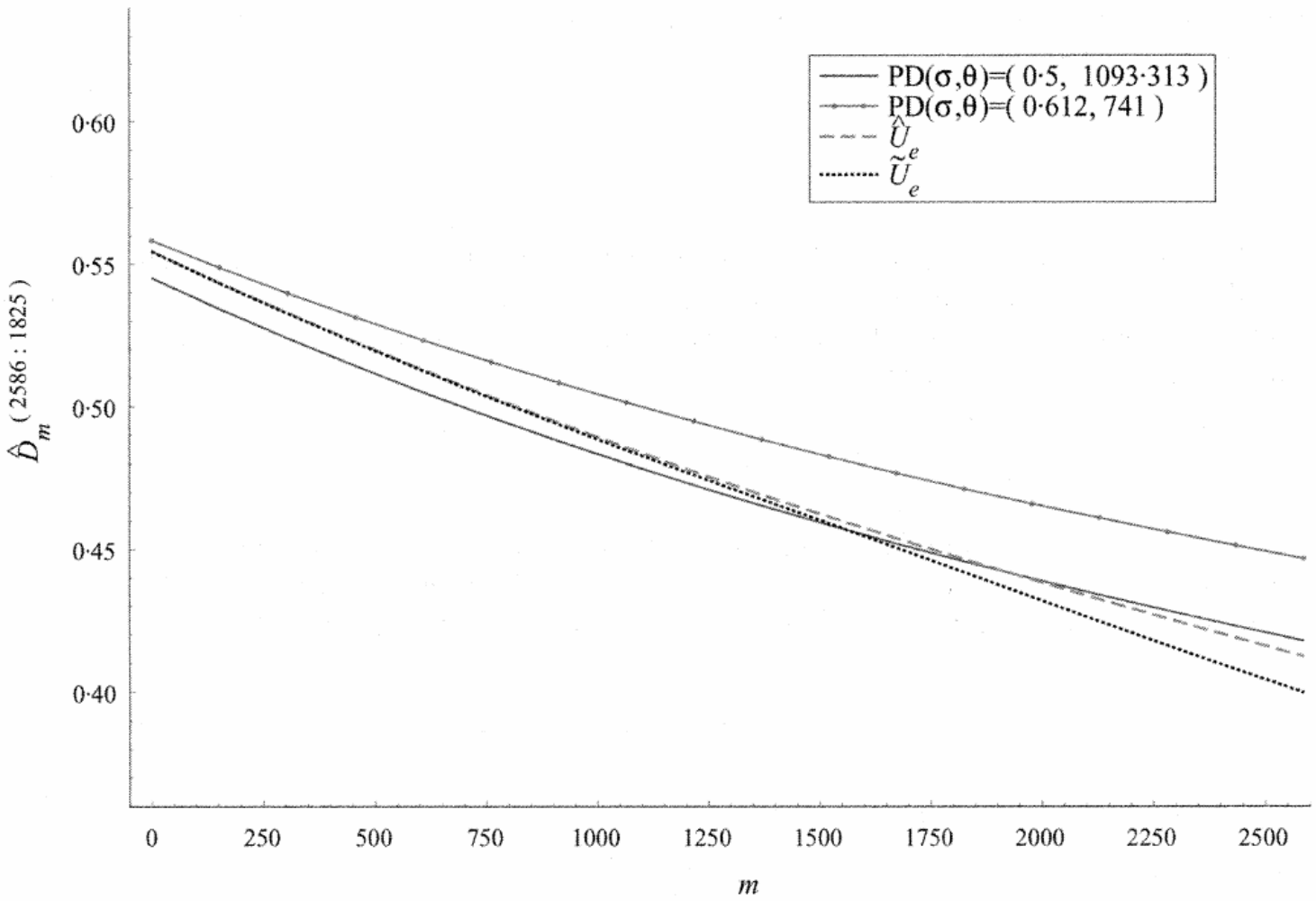

Fig. 4. Genomics example. Decay of the estimate $\hat{D}_m^{(2586:1825)}$ as m increases corresponding to the two choices of the Poisson–Dirichlet process, the moment-based estimator $\hat{U}_e$ and the likelihood-based estimator $\tilde{U}_e$.

References

Gnedin, A., and Pitman, J.
Exchangeable gibbs partitions and stirling triangles.
Journal of Mathematical sciences 138, 3 (2006), 5674–5685.

Lijoi, A., Mena, R. H., and Prünster, I.
Hierarchical mixture modeling with normalized inverse-gaussian priors.
Journal of the American Statistical Association 100, 472 (2005), 1278–1291.

Lijoi, A., Mena, R. H., and Prünster, I.
Bayesian nonparametric estimation of the probability of discovering new species.
Biometrika 94, 4 (2007), 769–786.

Mao, C. X.
Predicting the conditional probability of discovering a new class.
Journal of the American Statistical Association 99, 468 (2004), 1108–1118.

Pitman, J.
Exchangeable and partially exchangeable random partitions.
Probability theory and related fields 102, 2 (1995), 145–158.

Pitman, J.
Some developments of the blackwell-macqueen urn scheme.
Lecture Notes-Monograph Series (1996), 245–267.