Simulations

Lorenzo Ghilotti

2023-06-21

```
library(ProductFormFA)
```

In this section, we illustrate the performance of the different proposed models under different scenarios. Through these simulations, we aim at highlighting the model-specific properties that we have discussed in the previous sections, both in terms of in-sample rarefaction curve and in terms of prediction of the number of unseen features in future samples. To this end, we distinguish between two broad classes of generating mechanisms, i.e. (i) bounded-features scenarios, where the number of features observable in the population is bounded, that is $\exists K^* > 0$ such that $\lim_{n\to\infty} K_n = K^*$ almost surely; (ii) unbounded-features scenarios, where the number of features observable in the population is unbounded, that is $\lim_{n\to\infty} K_n = \infty$ almost surely. We are going to discuss how the two examples of feature allocation models - *Mixtures of IBP* and *Mixtures of Beta-Bernoulli with N features* - are suitable for different scenarios. Remark: it is not possible to establish with certainty whether the true generating mechanism belongs to a bounded-features scenario or an unbounded-features one by looking at the data. This decision pertains to the analyst, based on expert knowledge of the problem, translated into proper modeling assumptions. Through the following simulations, we just want to show that, if the true generating mechanism were known, then some models are more suitable than others to fit the data. As a consequence, when the knowledge of the problem suggests one of the two scenarios, some models might be preferred with respect to others.

Bounded-features scenarios

For these scenarios, we consider 5 ecological species detection models, as in Chiu (2022). Within these settings, the individuals are geographical sites where species of animals are collected (each species is a feature which might be displayed by the individual). In each scenario, the total number of species is $H = 500$ and the species occurrence probabilities $(\pi_1, \dots, \pi_H)$ are determined. We compare the Gamma mixture of IBP, the mixture of Beta-Bernoulli with Poisson prior on N and the mixture of Beta-Bernoulli with negative binomial prior on N. For each setting and each model we show the following quantities, estimated for different dimensions of the training set: (i.a) the in-sample rarefaction curve (on a single dataset), (i.b) the extrapolated rarefaction curve on the test set (on a single dataset), (ii) the accuracy of the estimated number of unseen features in the test sample, over $D = 50$ replicated datasets. Specifically, we focus on the following measure of accuracy, denoted with $\nu_m^{(n)}$,

$$\nu_m^{(n)} := \frac{1}{1 + \frac{|\tilde{K}_m^{(n)} - \hat{K}_m^{(n)}|}{\tilde{K}_n}},$$

where $\tilde{K}_m^{(n)}$ is the observed number of unseen features in the test set, $\hat{K}_m^{(n)}$ is the expected value of the statistic $K_m^{(n)}$ and $\tilde{K}_n$ is the observed number of features in the training set. Note that $\nu_m^{(n)} \in (0, 1]$, with $\nu_m^{(n)} = 1$ meaning perfect estimation.

Moreover, for the mixtures of Beta-Bernoulli, we also report (iii) the posterior distribution of the total number of features (on a single dataset), (iv) the expected value of the posterior distribution of the total number of features, over $D = 50$ replicated datasets.

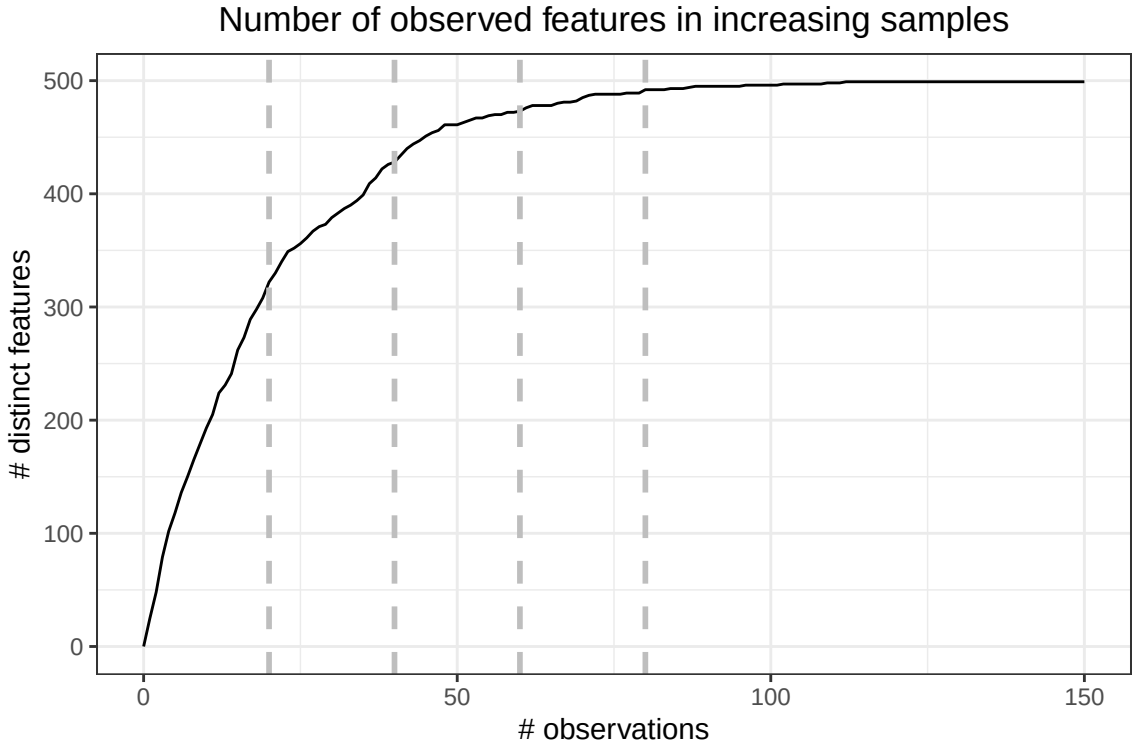
Scenario 1: the homogeneous model

Set $\pi_k = 0.05$, for $k = 1, \ldots, H$. Let L be the total dimension of the dataset, and consider different dimensions for the training set n, i.e.

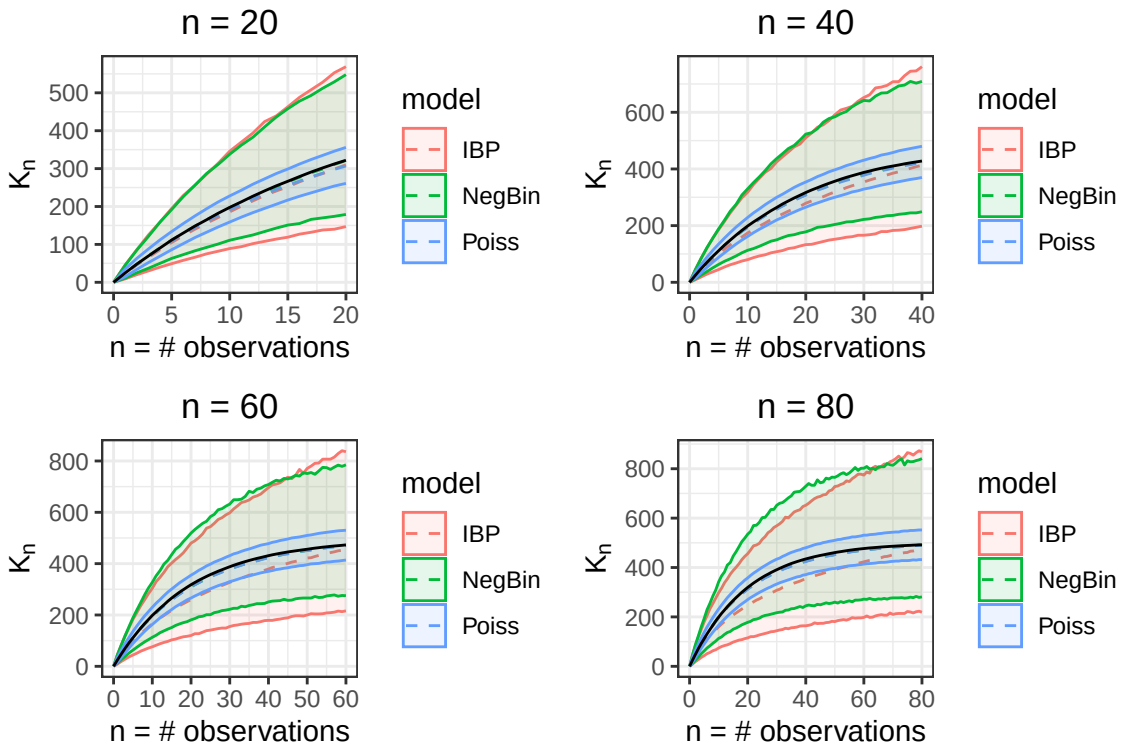
```
## L =  150
```

```
## n =  20  40  60  80
```

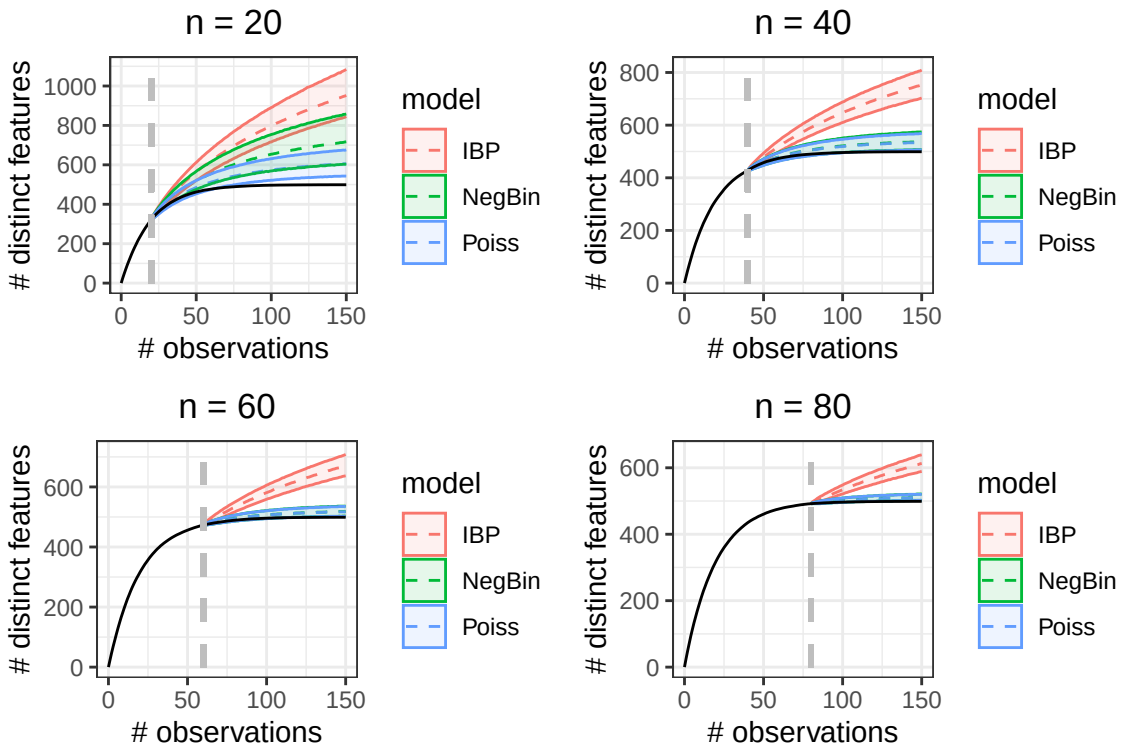
Here, the curve representing the number of observed features in increasing samples, where the grey vertical lines indicate the different training dimensions.



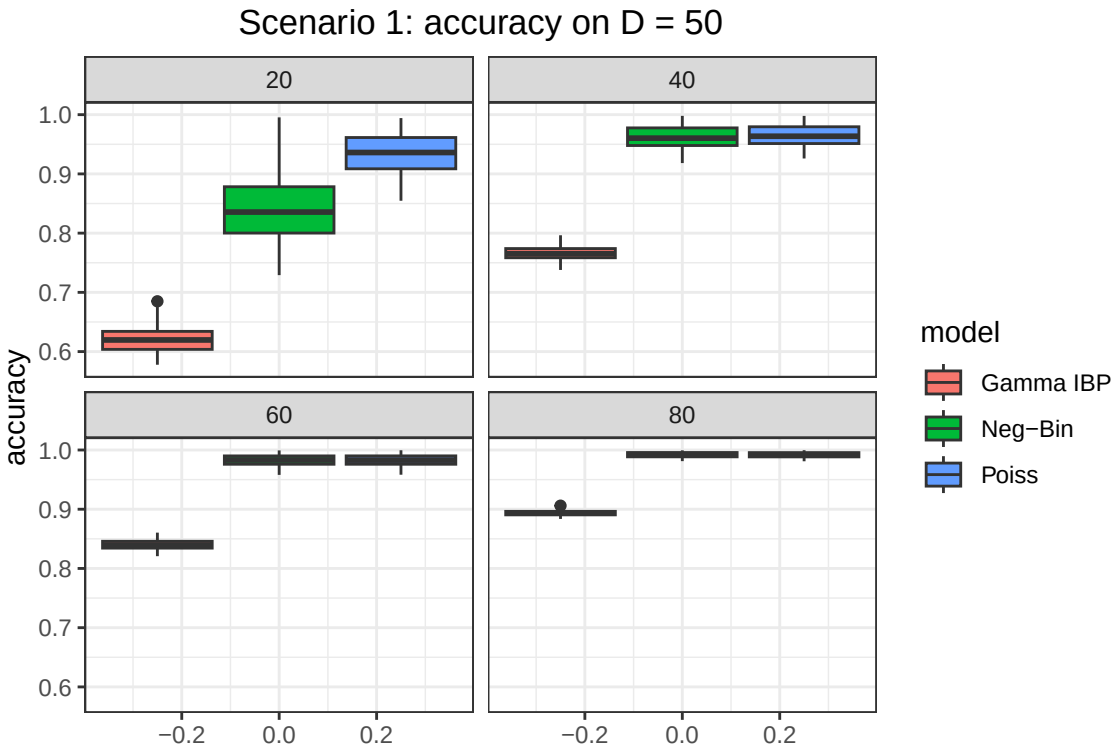
Here, we report (i.a) the in-sample rarefaction curve (on a single dataset).



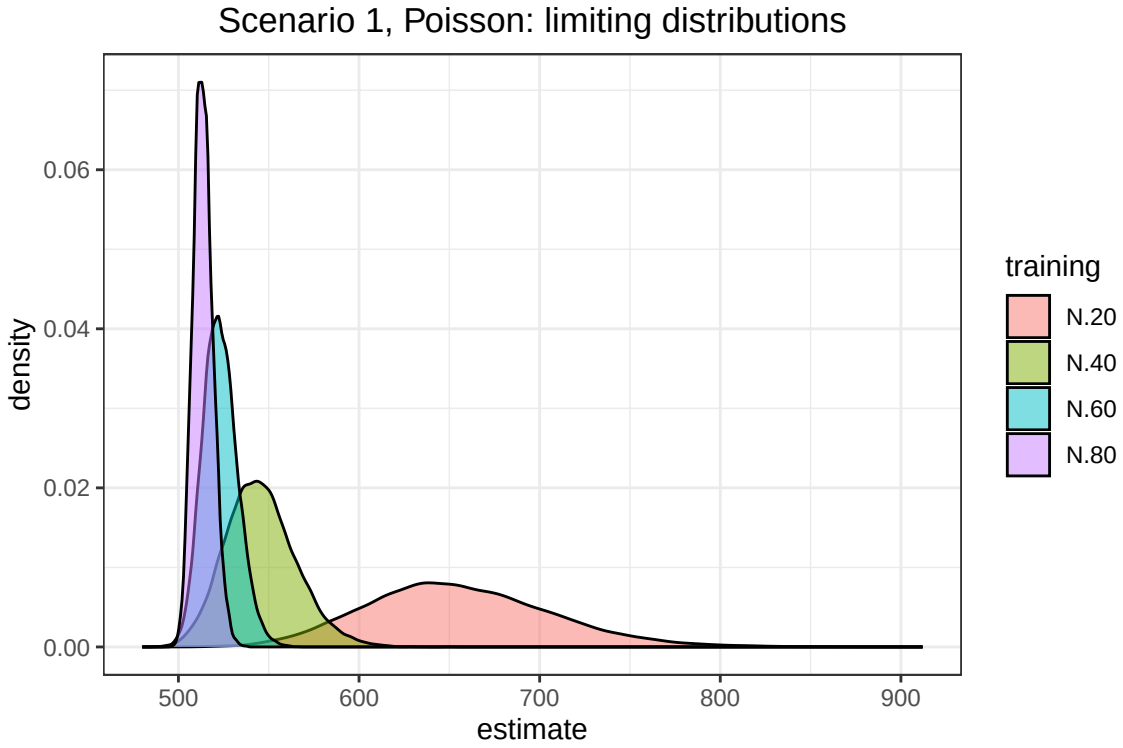
Here, we report (i.b) the extrapolated rarefaction curve (on a single dataset).

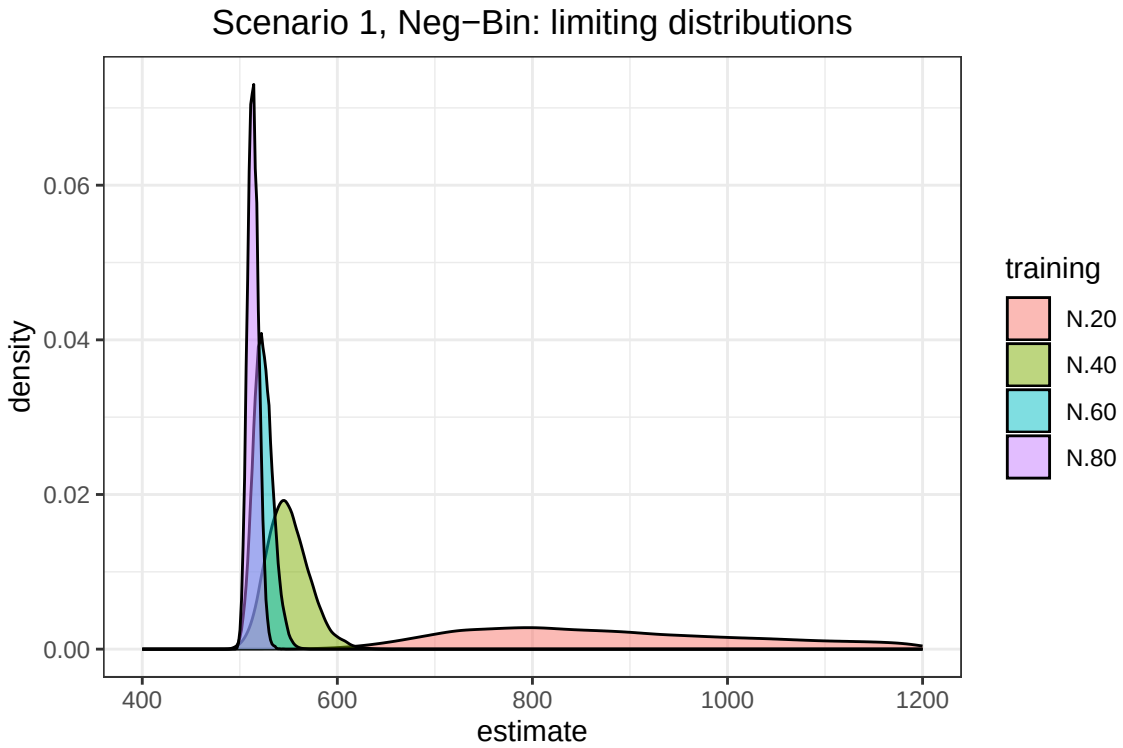


Here, we report (ii) the accuracy of the estimated number of unseen features in the test sample, over $D = 50$ replicated datasets.

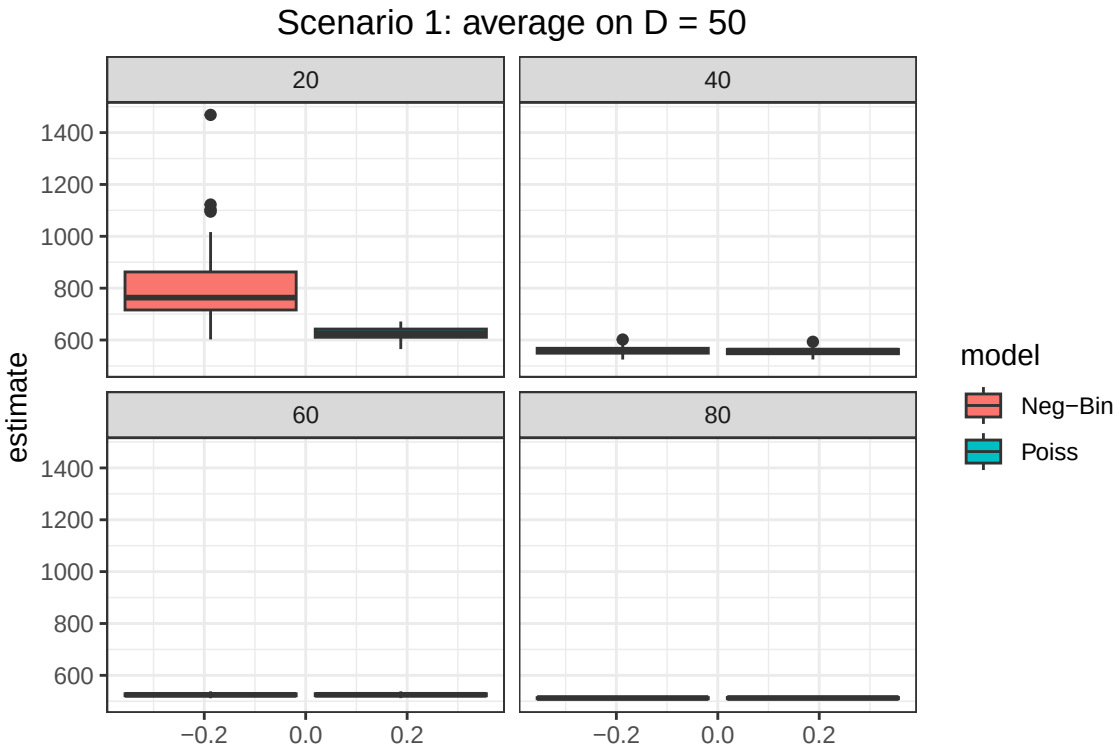


For the mixtures of Beta-Bernoulli, we report (iii) the posterior distribution of the total number of features (on a single dataset).





Finally, we report (iv) the expected value of the posterior distribution of the total number of features, over $D = 50$ replicated datasets.



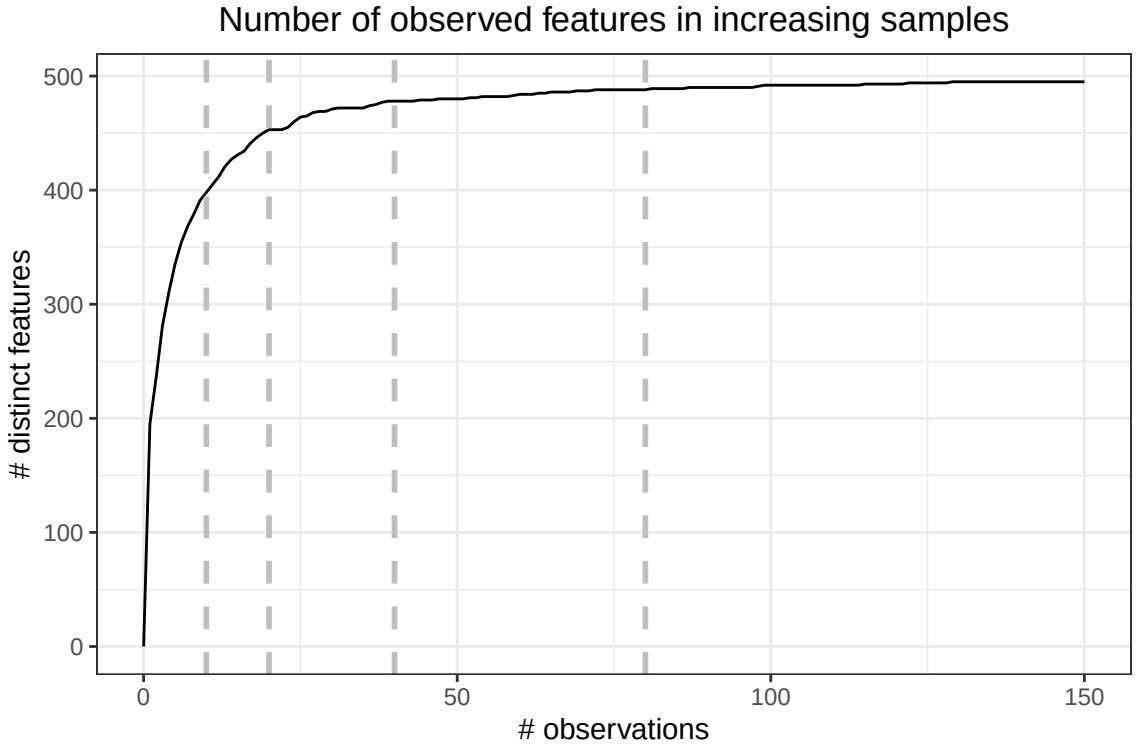
Scenario 2: the random uniform model

Set $\pi_k = c \cdot a_k$, for $k = 1, \ldots, H$, and $a_k \overset{iid}{\sim} \text{Uniform}(0, 1)$. Set c such that the maximum π_k is equal to 0.5. Let L be the total dimension of the dataset, and consider different dimensions for the training set n, i.e.

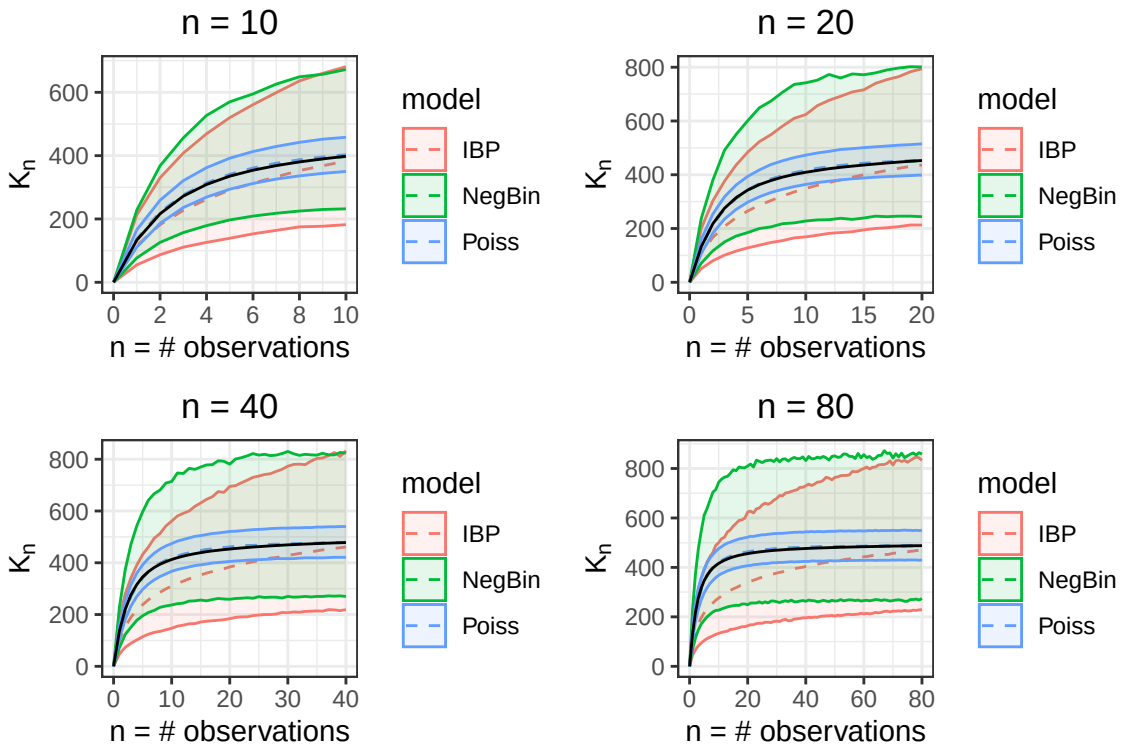
```
## L =  150
```

```
## n =  10  20  40  80
```

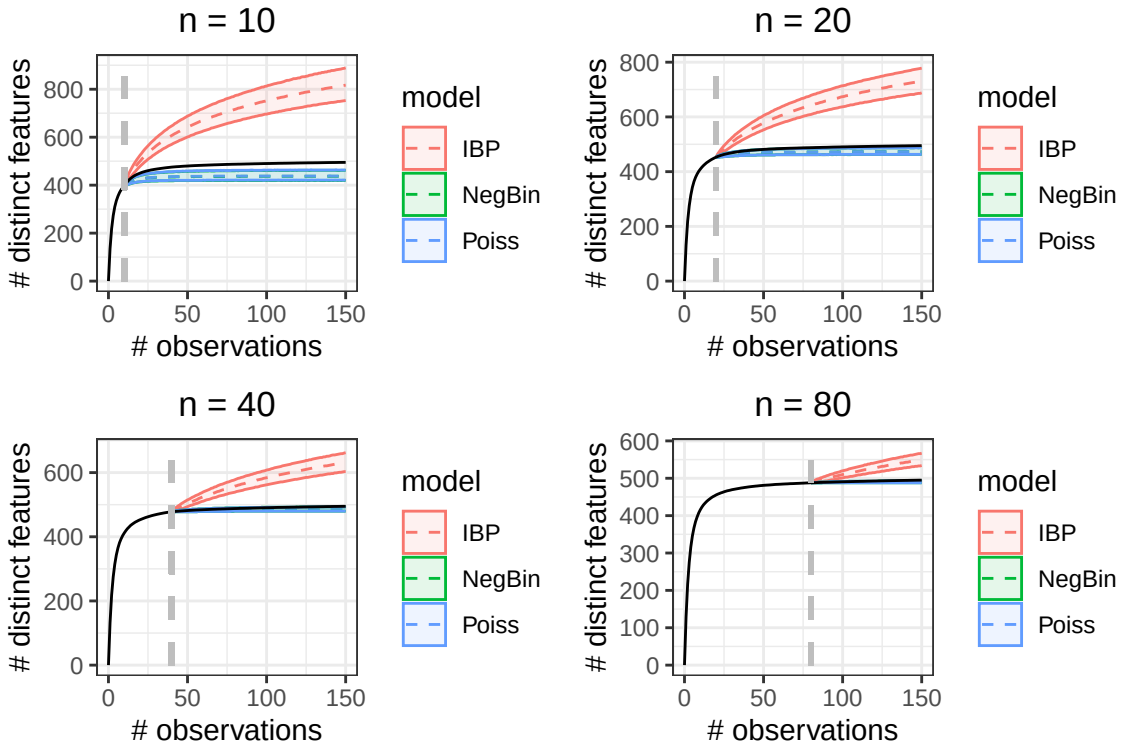
Here, the curve representing the number of observed features in increasing samples, where the grey vertical lines indicate the different training dimensions.



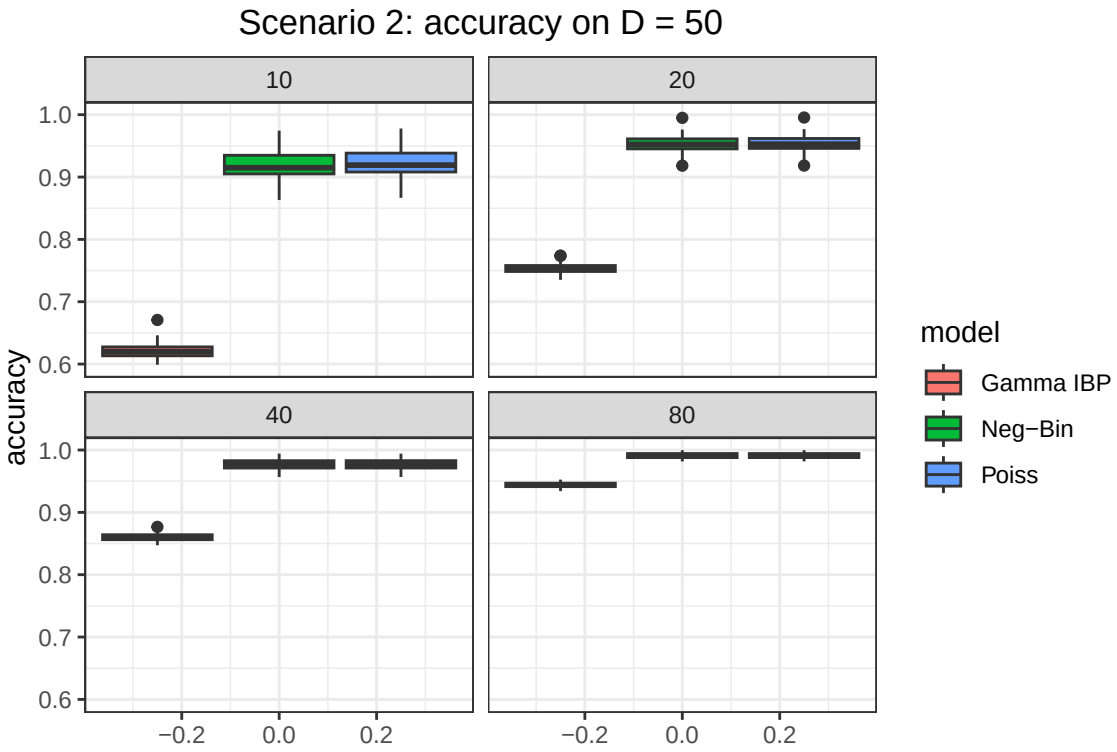
Here, we report (i.a) the in-sample rarefaction curve (on a single dataset).



Here, we report (i.b) the extrapolated rarefaction curve (on a single dataset).

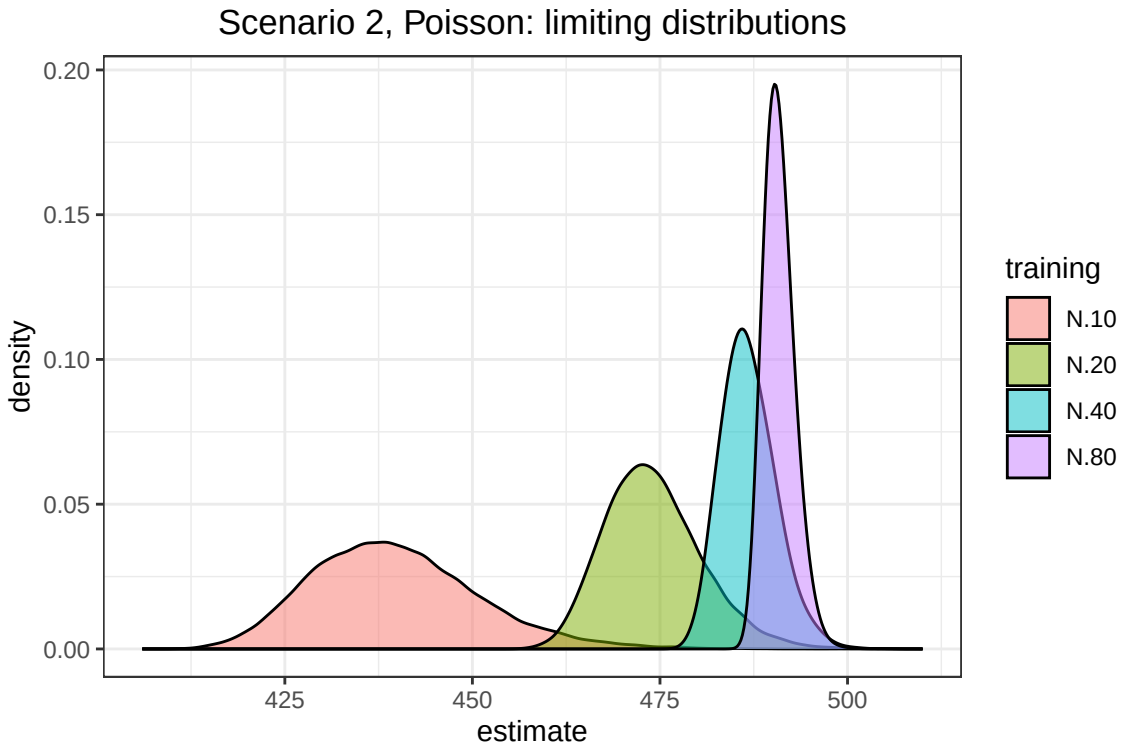


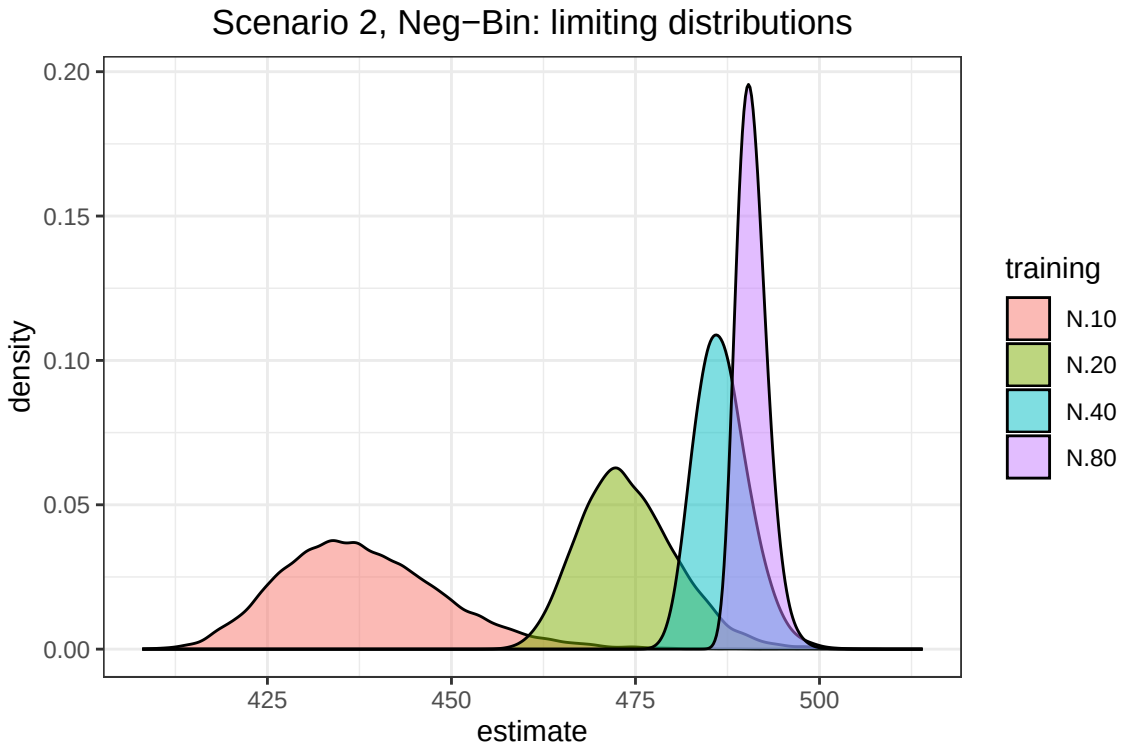
Here, we report (ii) the accuracy of the estimated number of unseen features in the test sample, over $D = 50$ replicated datasets.



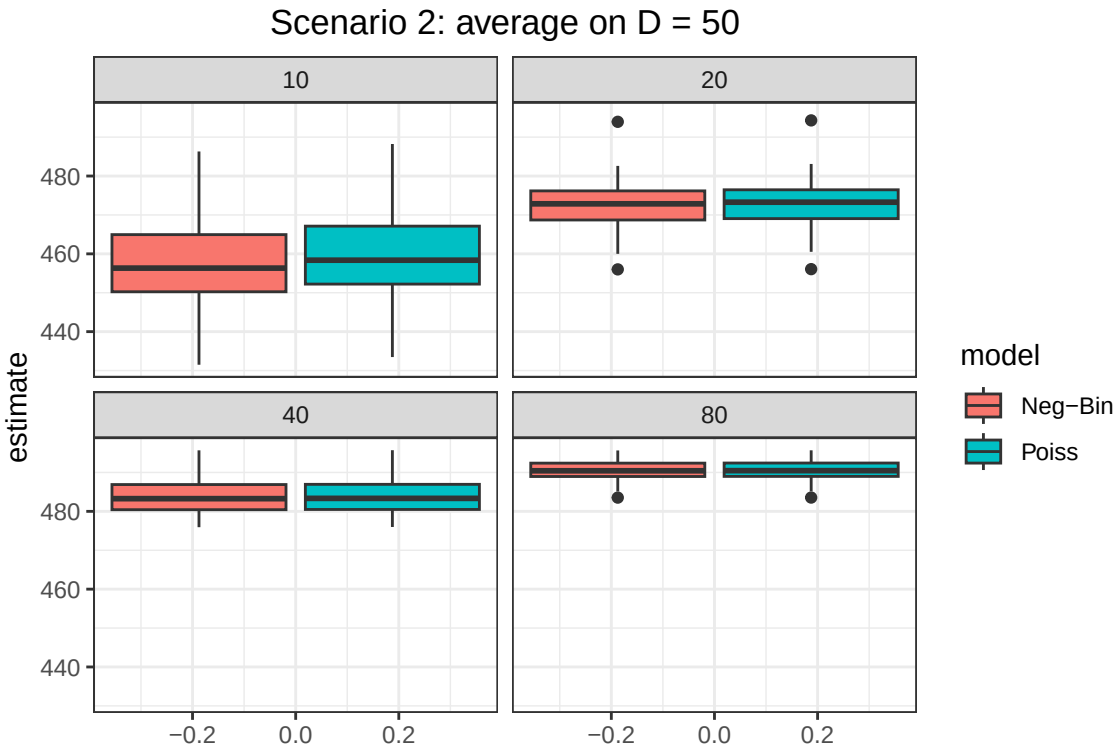
Even if the Mixture of IBP seems to reach better performance when the training set increases, this is just due to the fact the test set dimension is reducing: see the behaviour of the extrapolated rarefaction curve to get the behaviour of the model on larger test sets.

For the mixtures of Beta-Bernoulli, we report (iii) the posterior distribution of the total number of features (on a single dataset).





Finally, we report (iv) the expected value of the posterior distribution of the total number of features, over $D = 50$ replicated datasets.



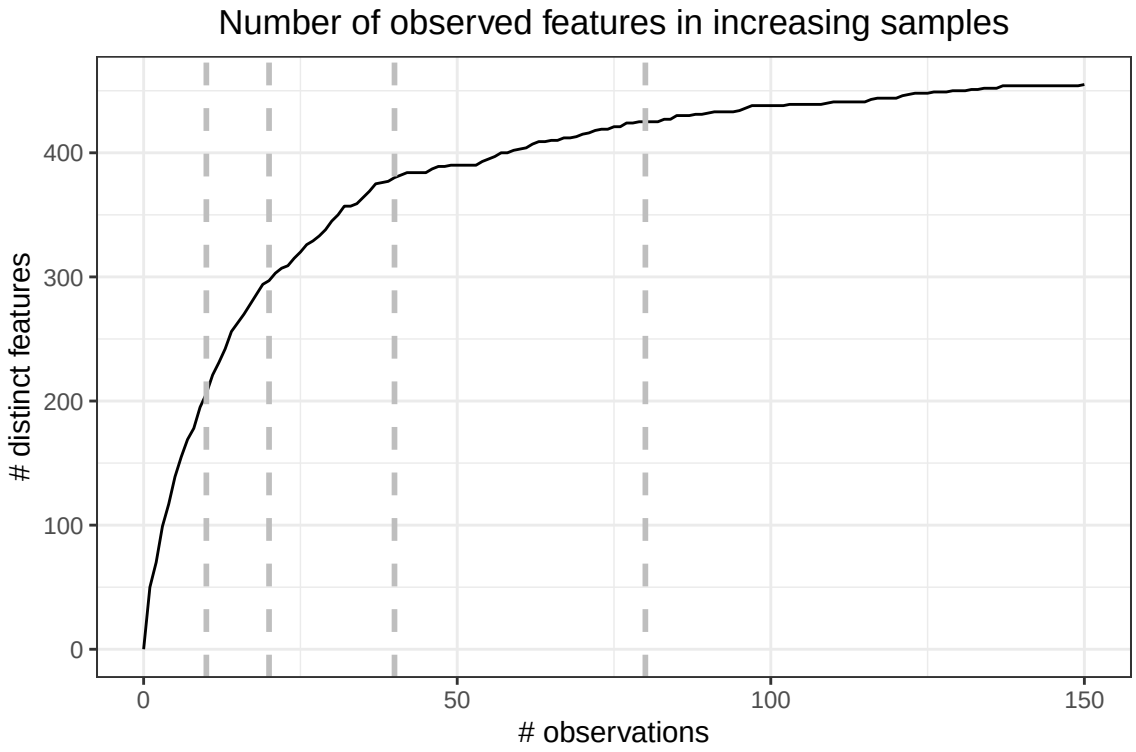
Scenario 3: the broken stick model

Set $\pi_k = c \cdot a_k$, for $k = 1, \ldots, H$, and $a_k \overset{iid}{\sim} \text{Exp}(1)$. Set c such that the maximum π_k is equal to 0.5. As Chiu (2022) says: "This model is commonly used in previous literature and equivalent to the Dirichlet distribution". Let L be the total dimension of the dataset, and consider different dimensions for the training set n, i.e.

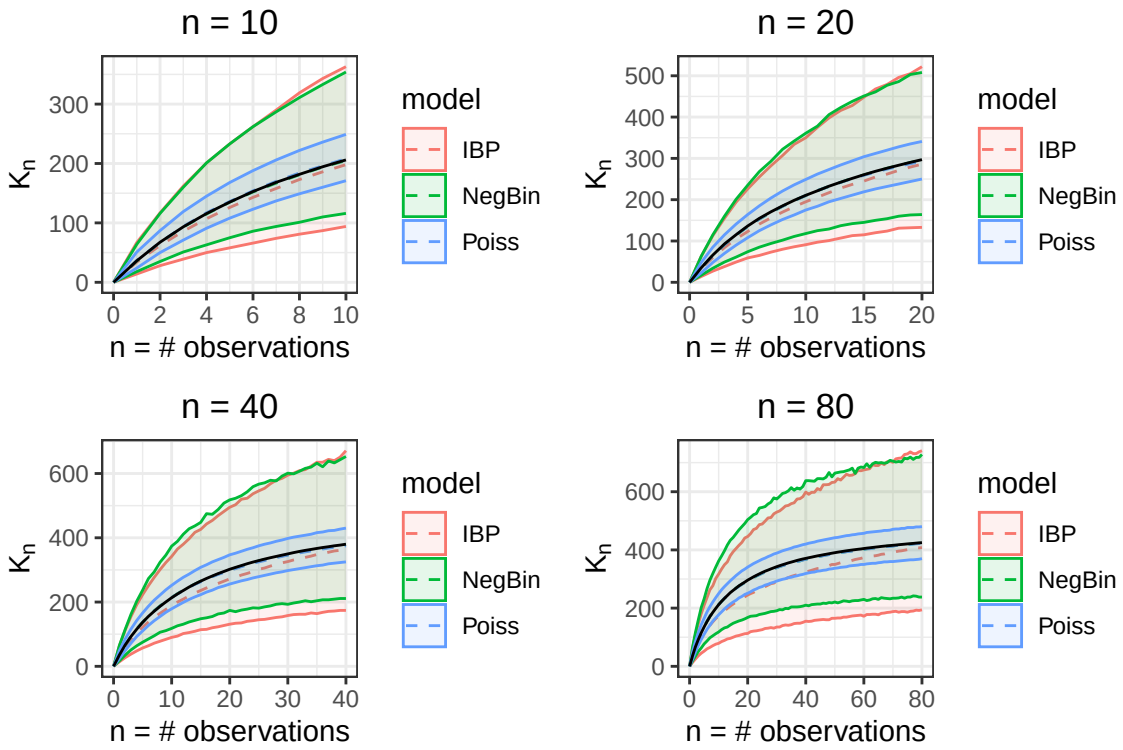
```
## L =  150
```

```
## n =  10  20  40  80
```

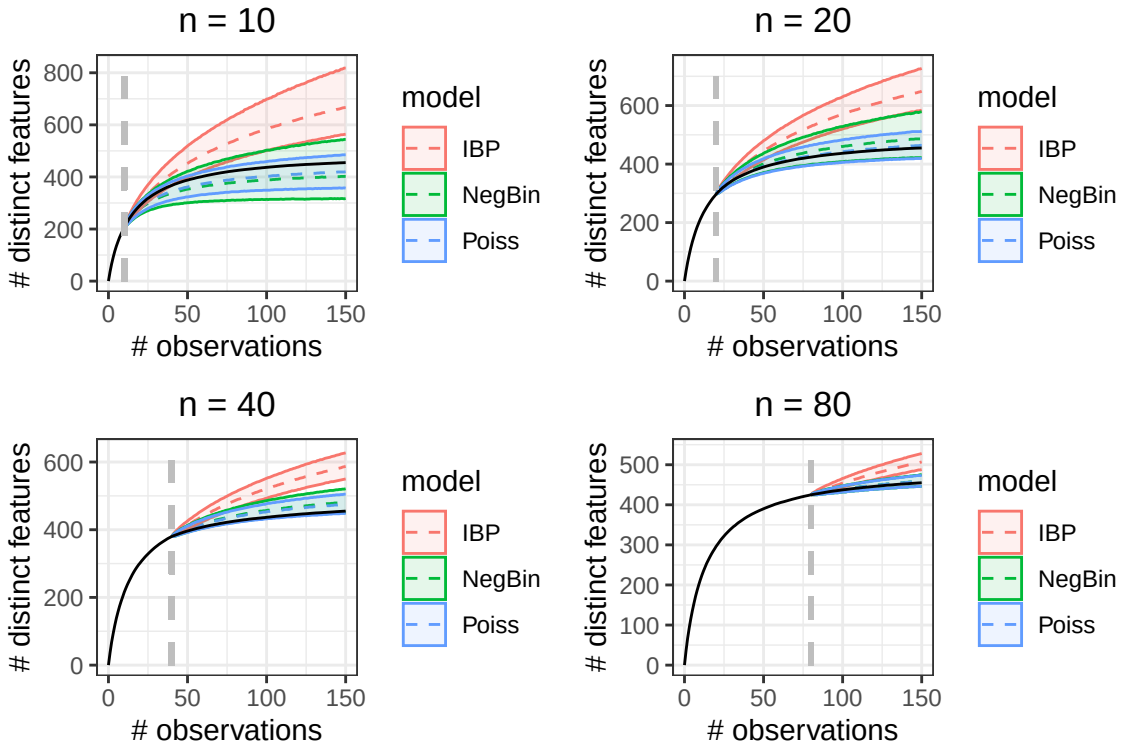
Here, the curve representing the number of observed features in increasing samples, where the grey vertical lines indicate the different training dimensions.



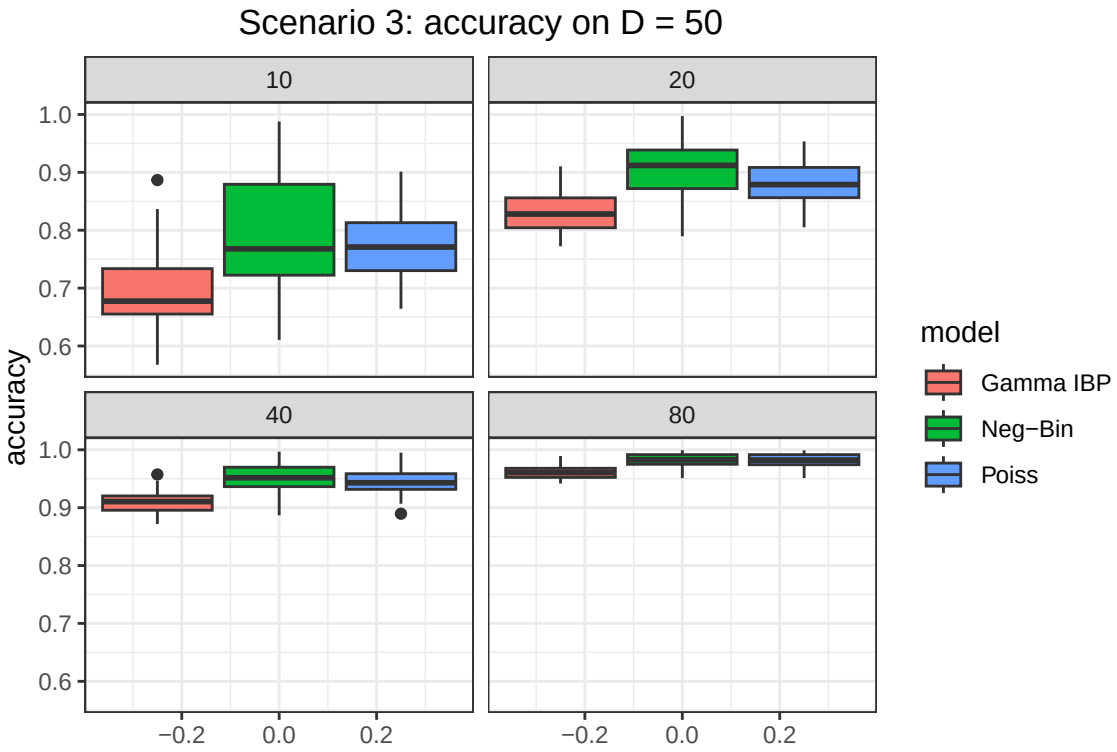
Here, we report (i.a) the in-sample rarefaction curve (on a single dataset).



Here, we report (i.b) the extrapolated rarefaction curve (on a single dataset).

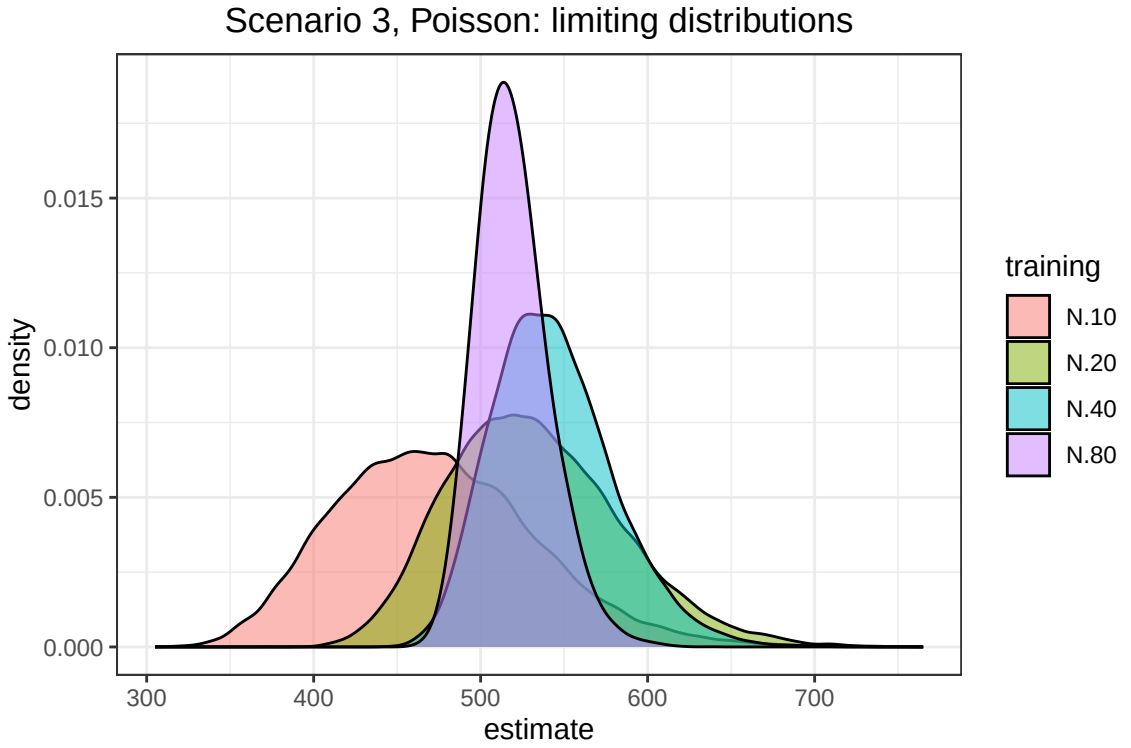


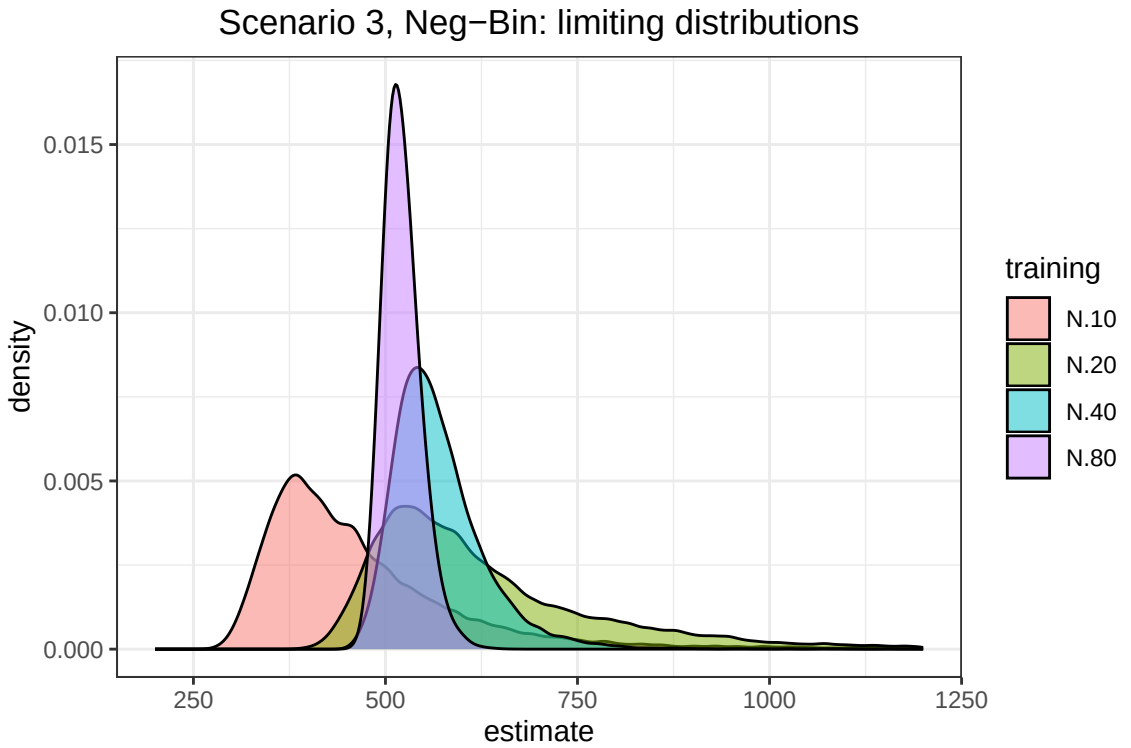
Here, we report (ii) the accuracy of the estimated number of unseen features in the test sample, over $D = 50$ replicated datasets.



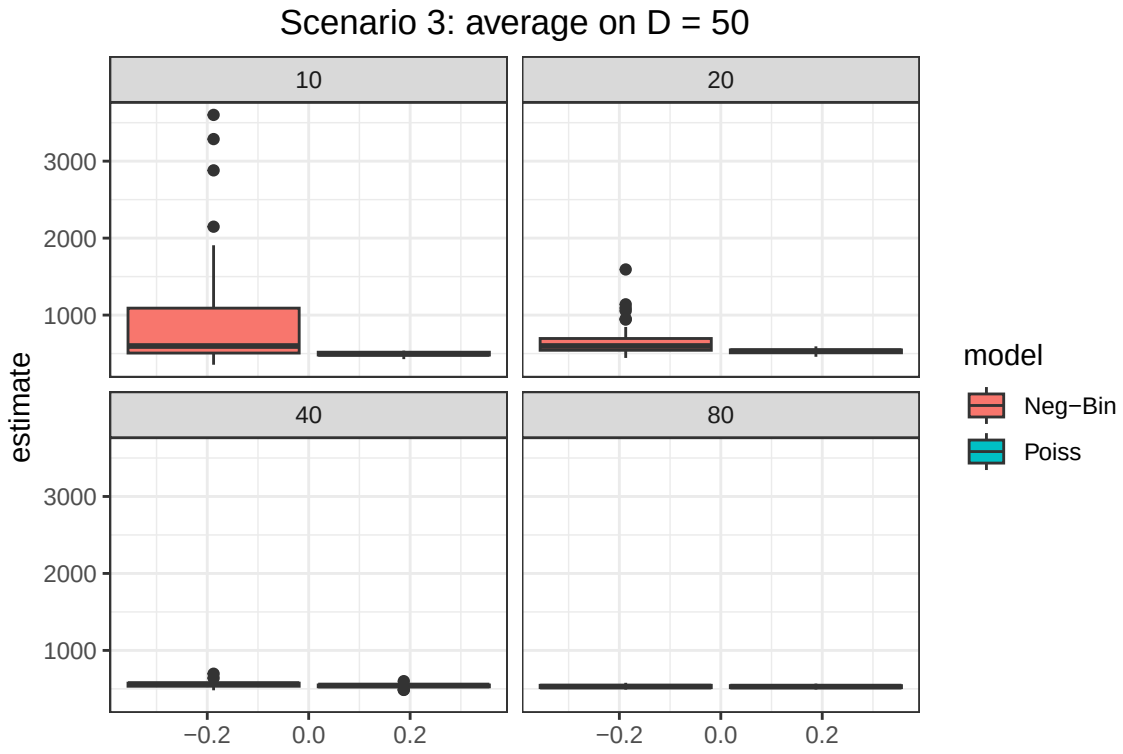
Even if the Mixture of IBP seems to reach better performance when the training set increases, this is just due to the fact the test set dimension is reducing: see the behavior of the extrapolated rarefaction curve to get the behavior of the model on larger test sets.

For the mixtures of Beta-Bernoulli, we report (iii) the posterior distribution of the total number of features (on a single dataset).





Finally, we report (iv) the expected value of the posterior distribution of the total number of features, over $D = 50$ replicated datasets.



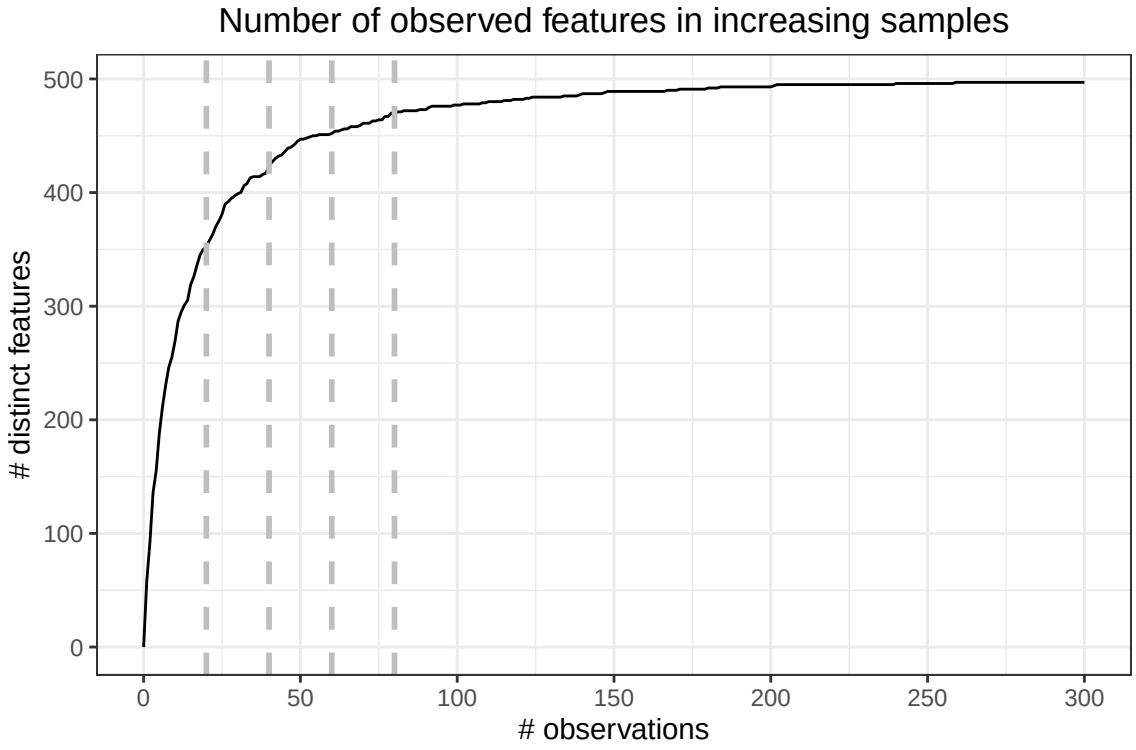
Scenario 4: the log-normal model

Set $\pi_k = c \cdot a_k$, for $k = 1, \ldots, H$, and $a_k \overset{iid}{\sim} \text{lognormal}(0, 1)$. Set c such that the maximum π_k is equal to 1. Let L be the total dimension of the dataset, and consider different dimensions for the training set n, i.e.

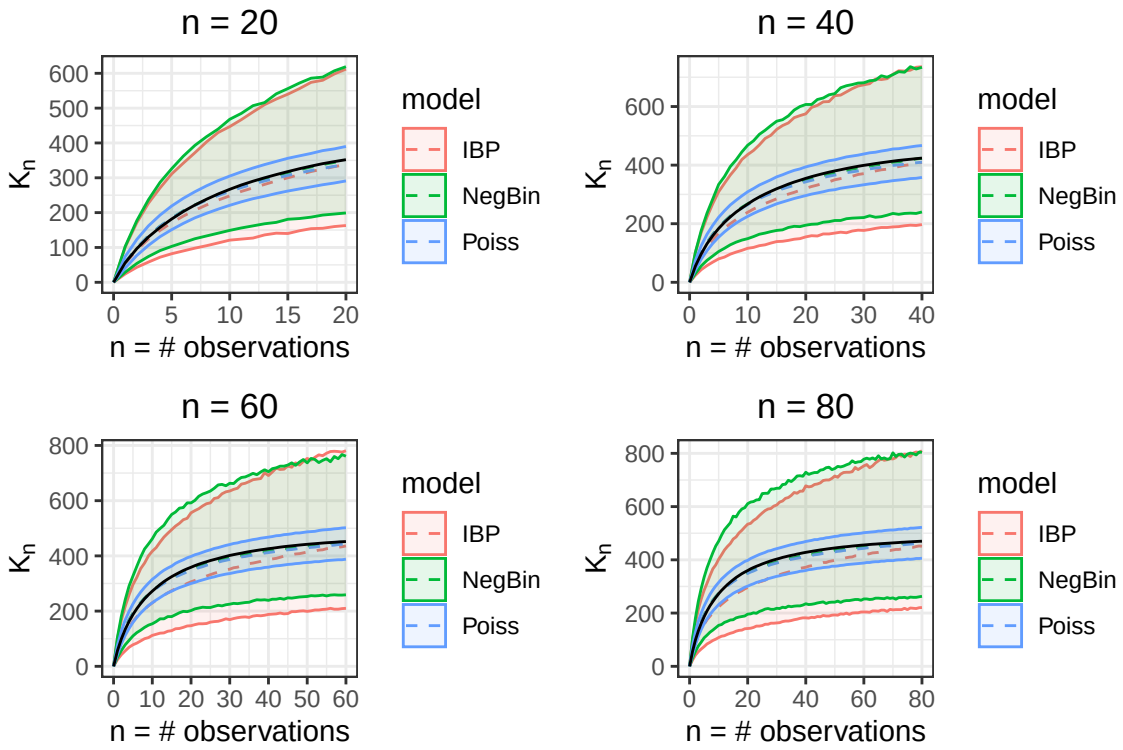
```
## L =  300
```

```
## n =  20  40  60  80
```

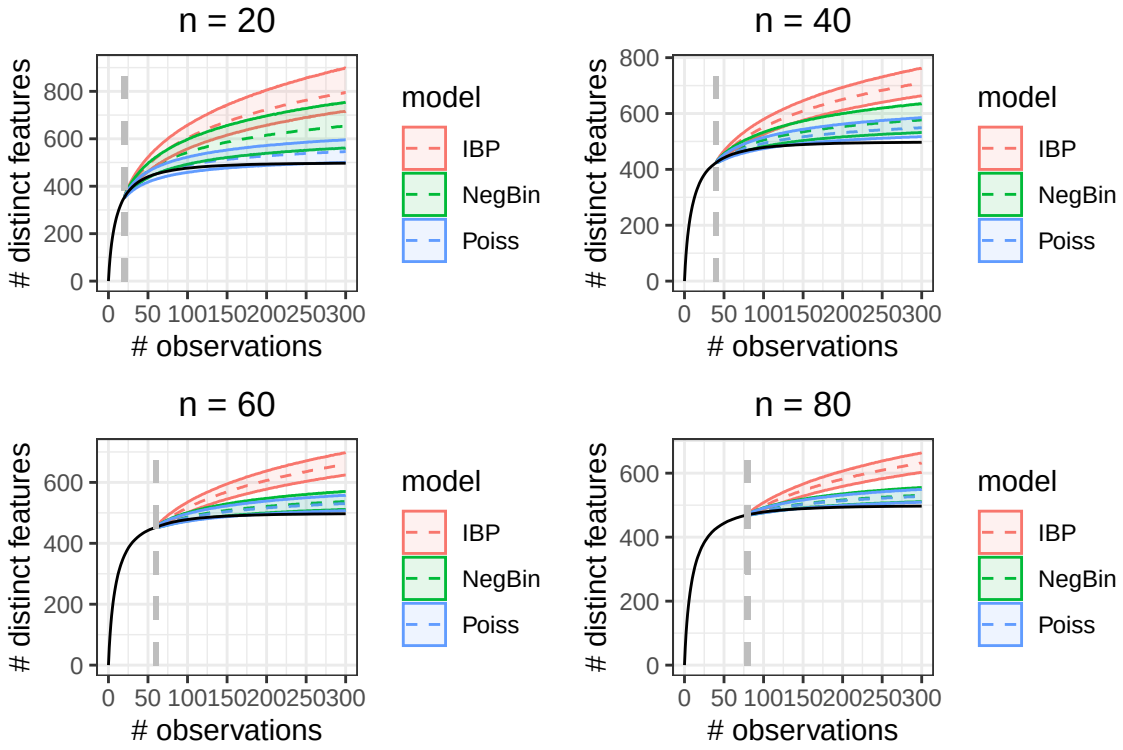
Here, the curve representing the number of observed features in increasing samples, where the grey vertical lines indicate the different training dimensions.



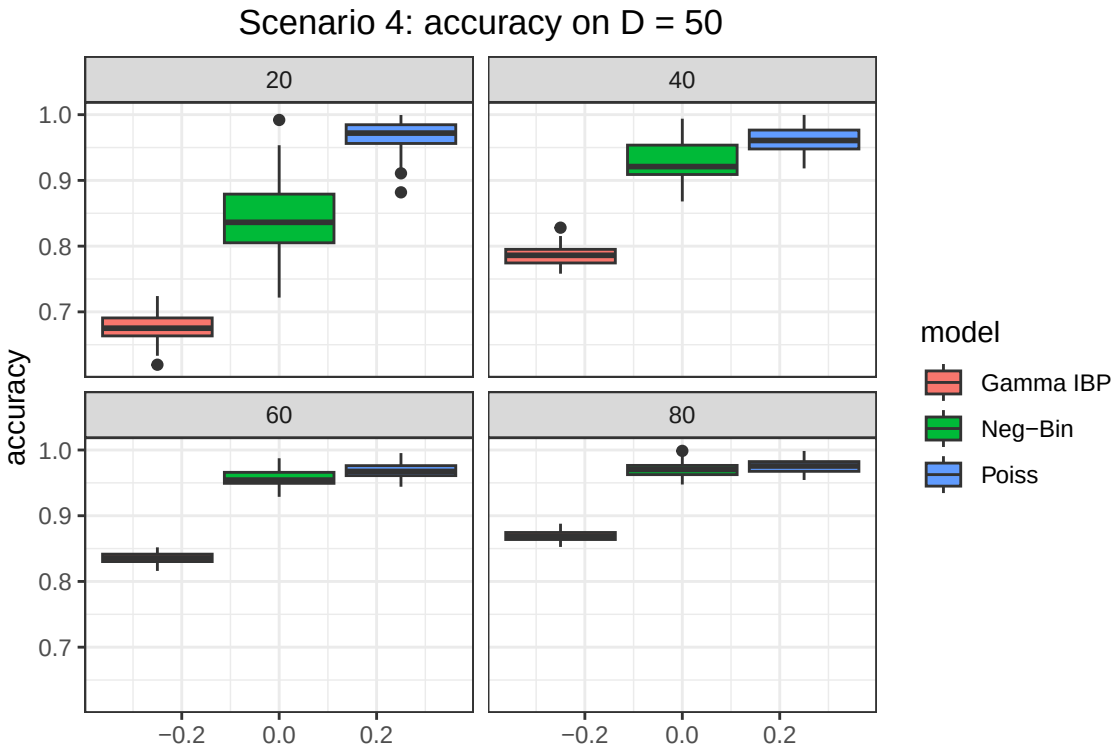
Here, we report (i.a) the in-sample rarefaction curve (on a single dataset).



Here, we report (i.b) the extrapolated rarefaction curve (on a single dataset).

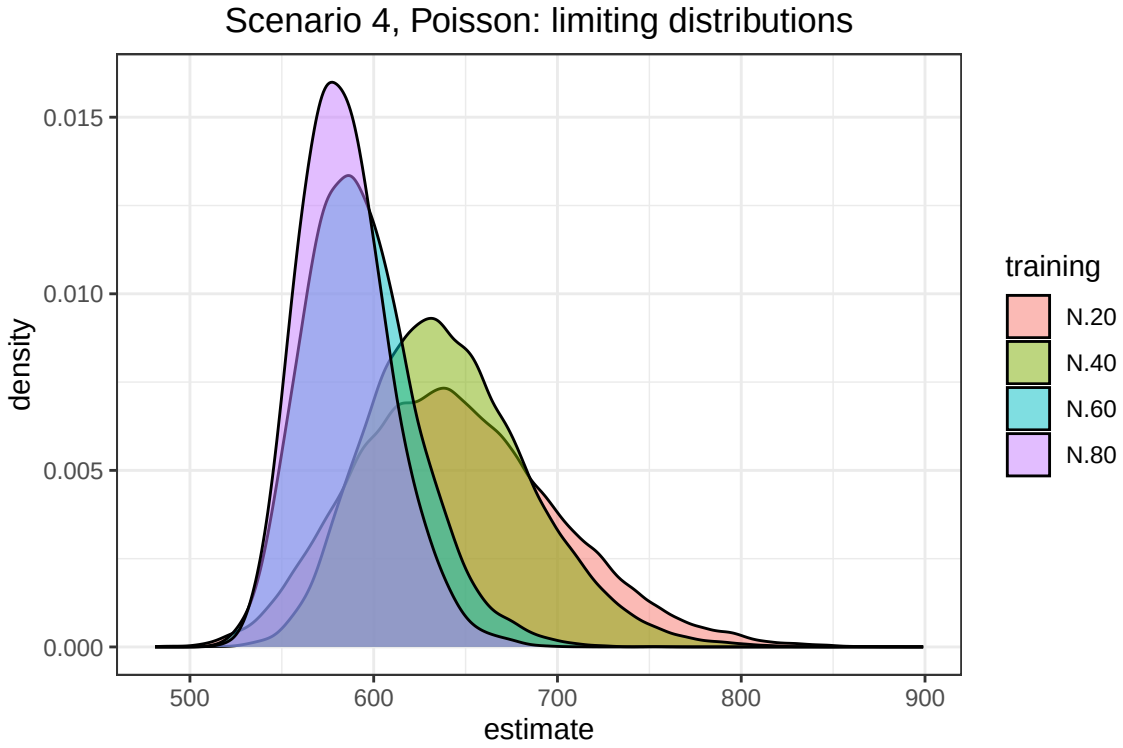


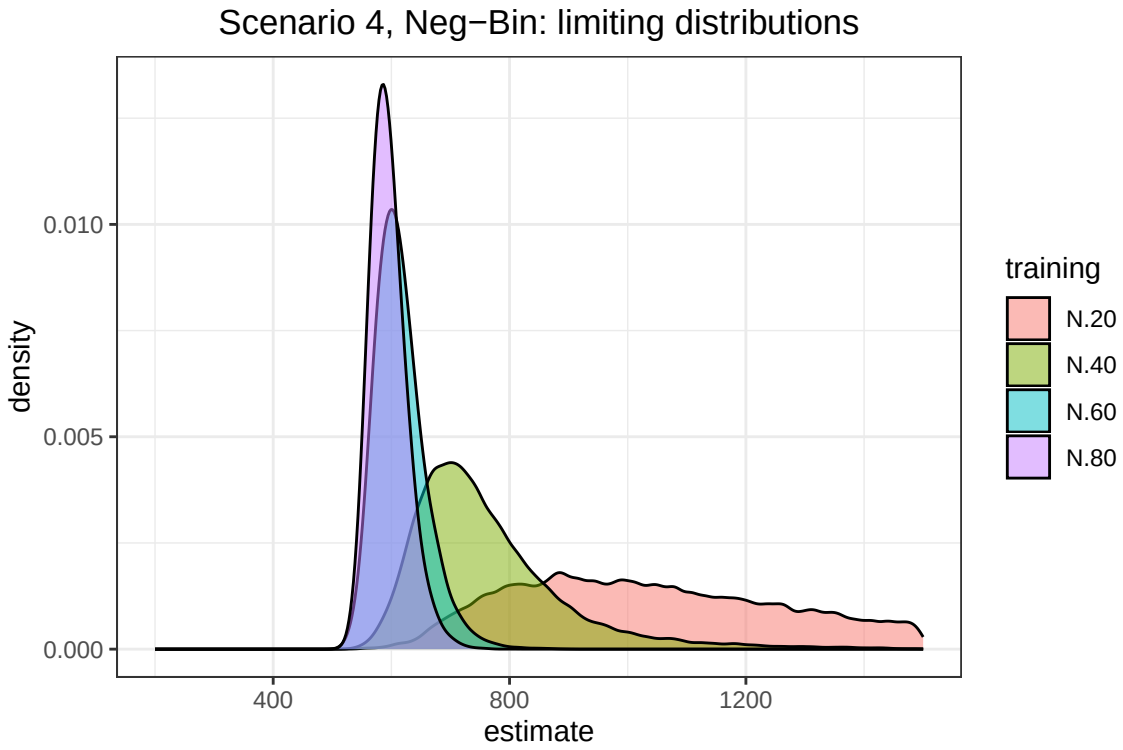
Here, we report (ii) the accuracy of the estimated number of unseen features in the test sample, over $D = 50$ replicated datasets.



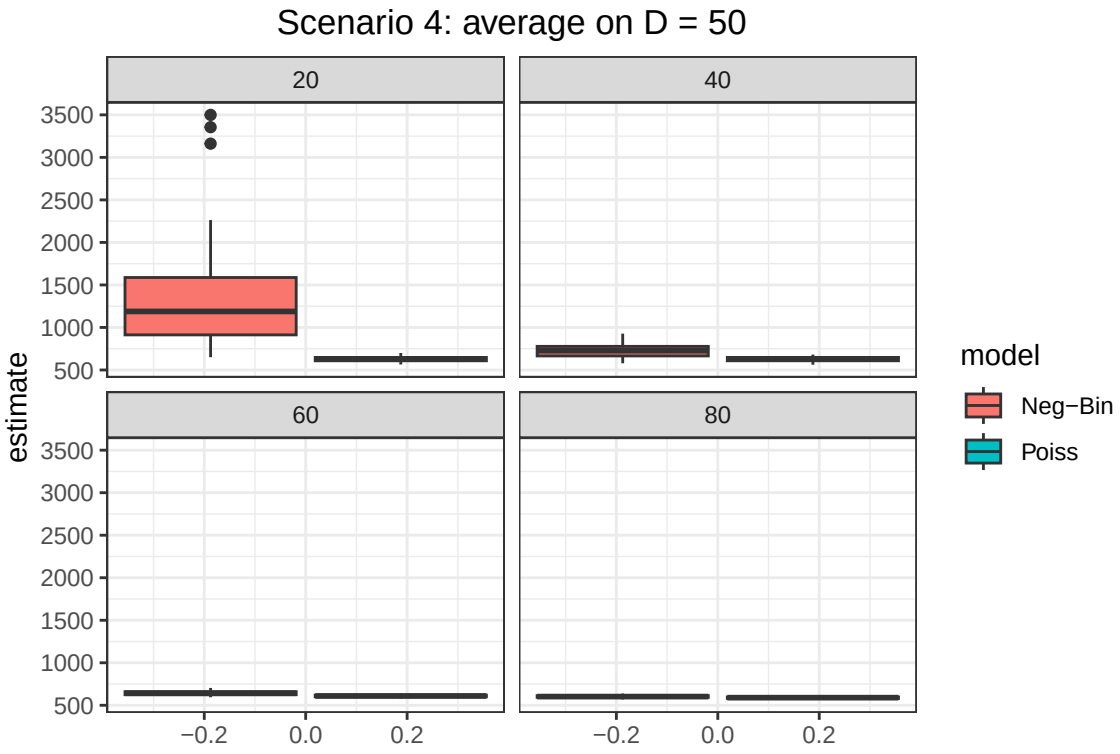
Even if the Mixture of IBP seems to reach better performance when the training set increases, this is just due to the fact the test set dimension is reducing: see the behavior of the extrapolated rarefaction curve to get the behavior of the model on larger test sets.

For the mixtures of Beta-Bernoulli, we report (iii) the posterior distribution of the total number of features (on a single dataset).





Finally, we report (iv) the expected value of the posterior distribution of the total number of features, over $D = 50$ replicated datasets.



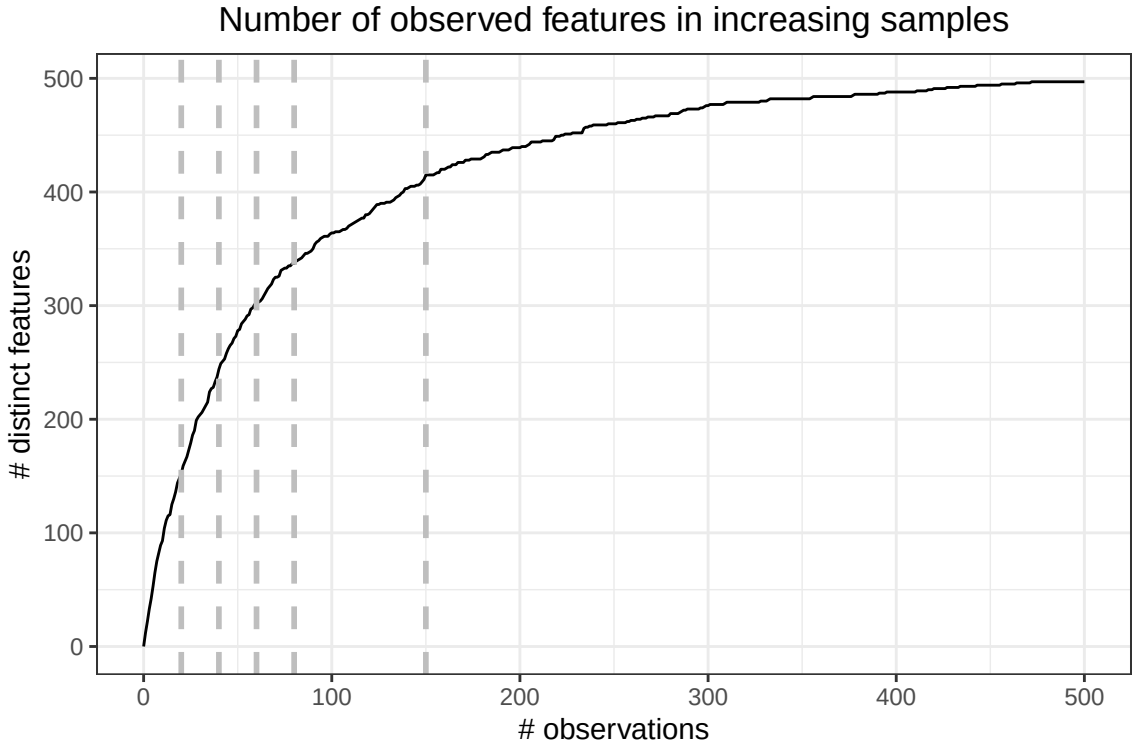
Scenario 5: the Zipf–Mandelbrot model

Set $\pi_k = \frac{3}{k+5}$, for $k = 1, \ldots, H$. Note that the maximum π_k is equal to 0.5. Let L be the total dimension of the dataset, and consider different dimensions for the training set n, i.e.

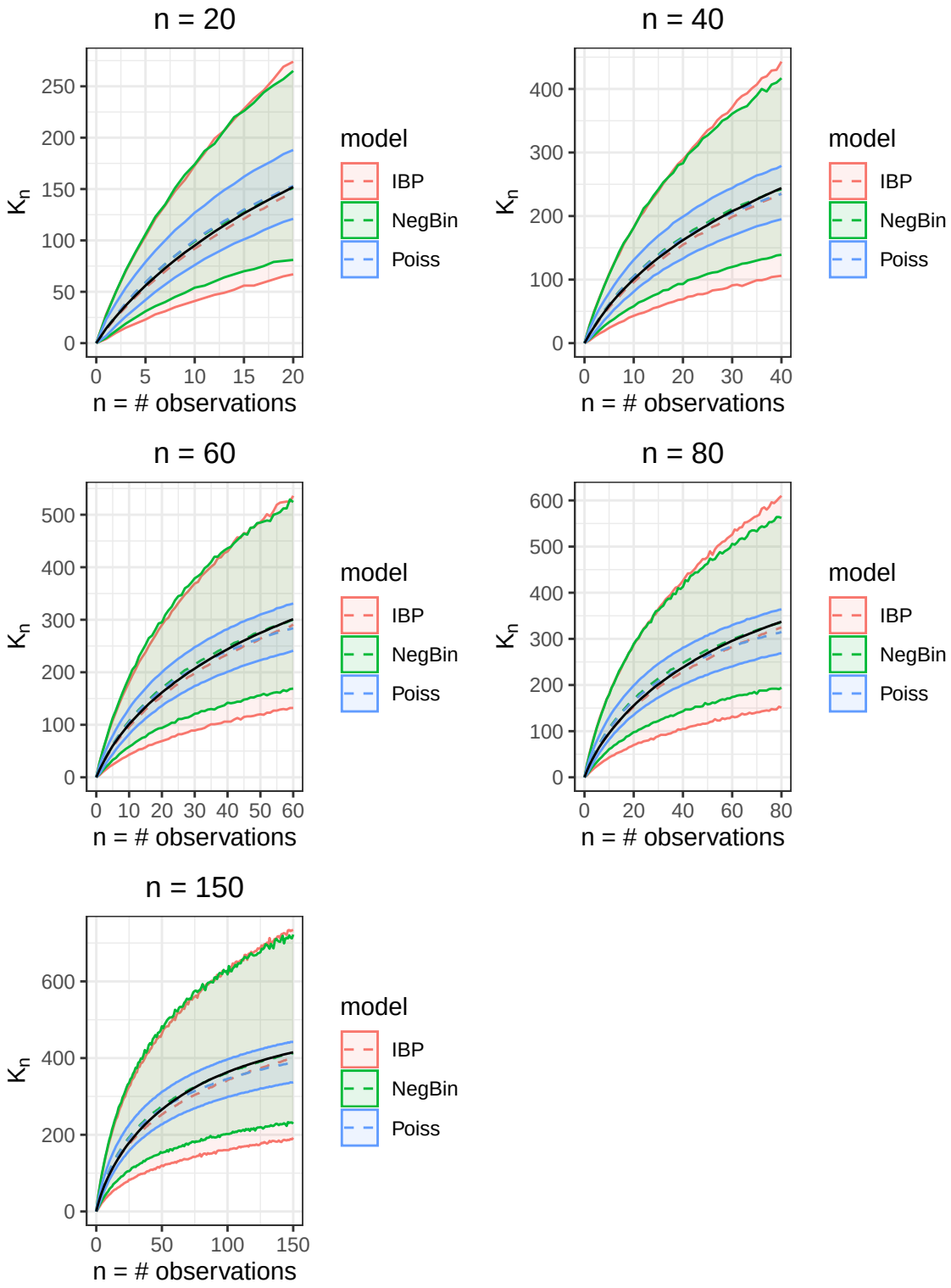
```
## L =  500
```

```
## n =  20  40  60  80  150
```

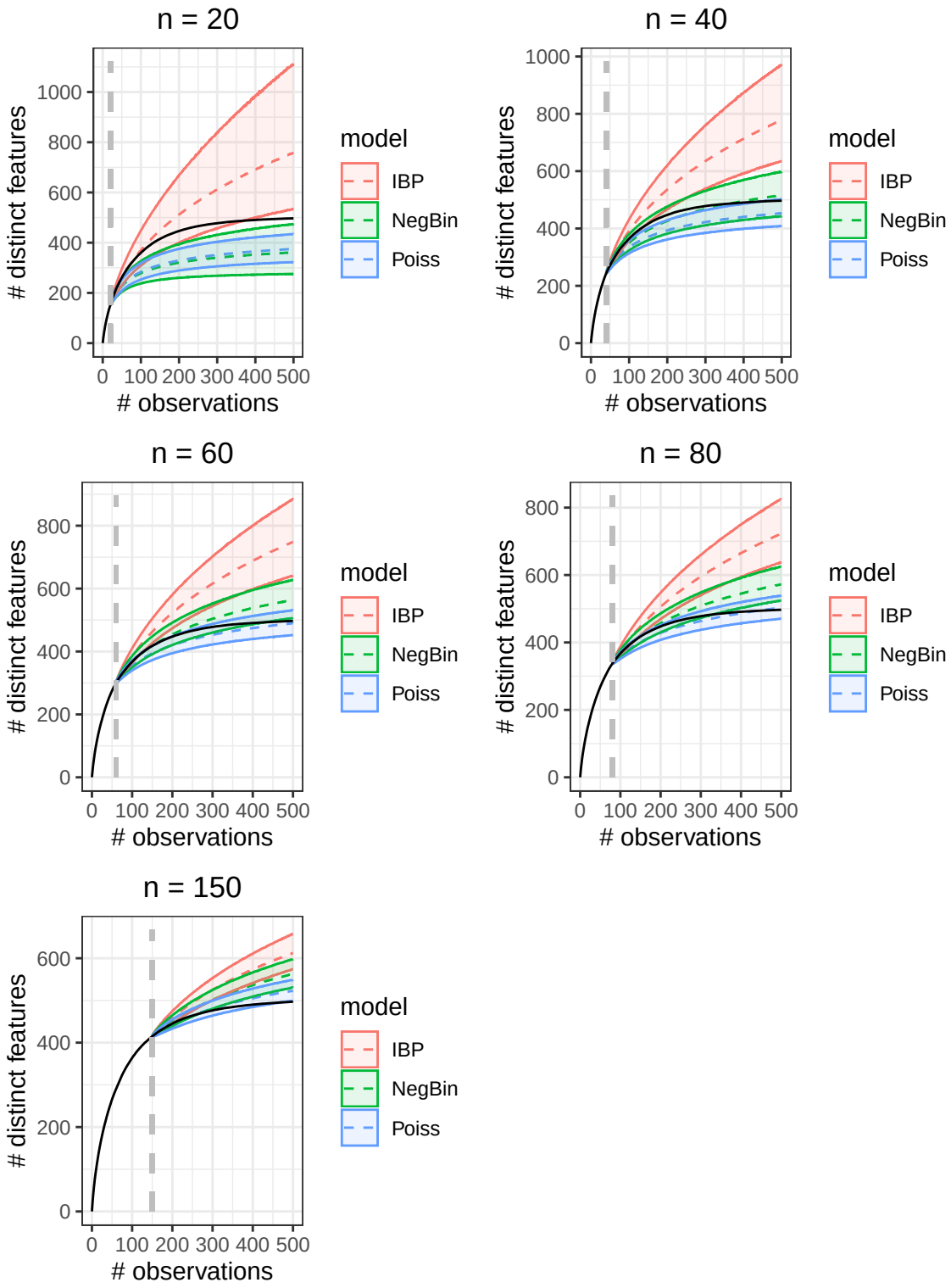
Here, the curve representing the number of observed features in increasing samples, where the grey vertical lines indicate the different training dimensions.



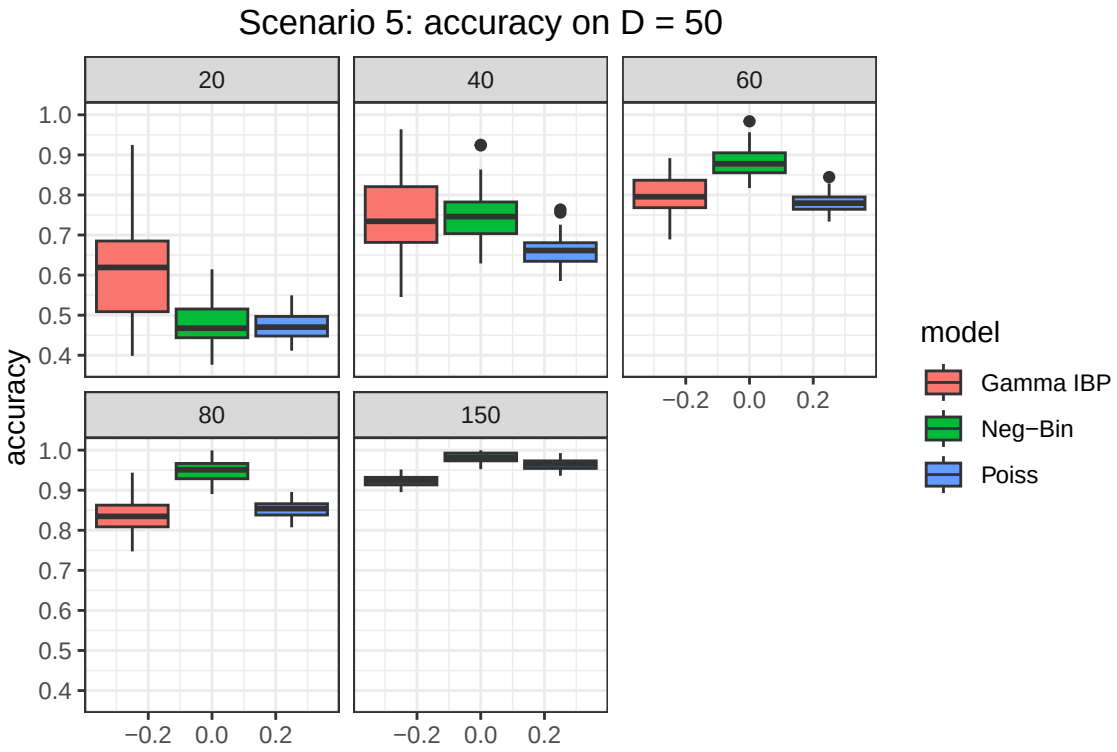
Here, we report (i.a) the in-sample rarefaction curve (on a single dataset).



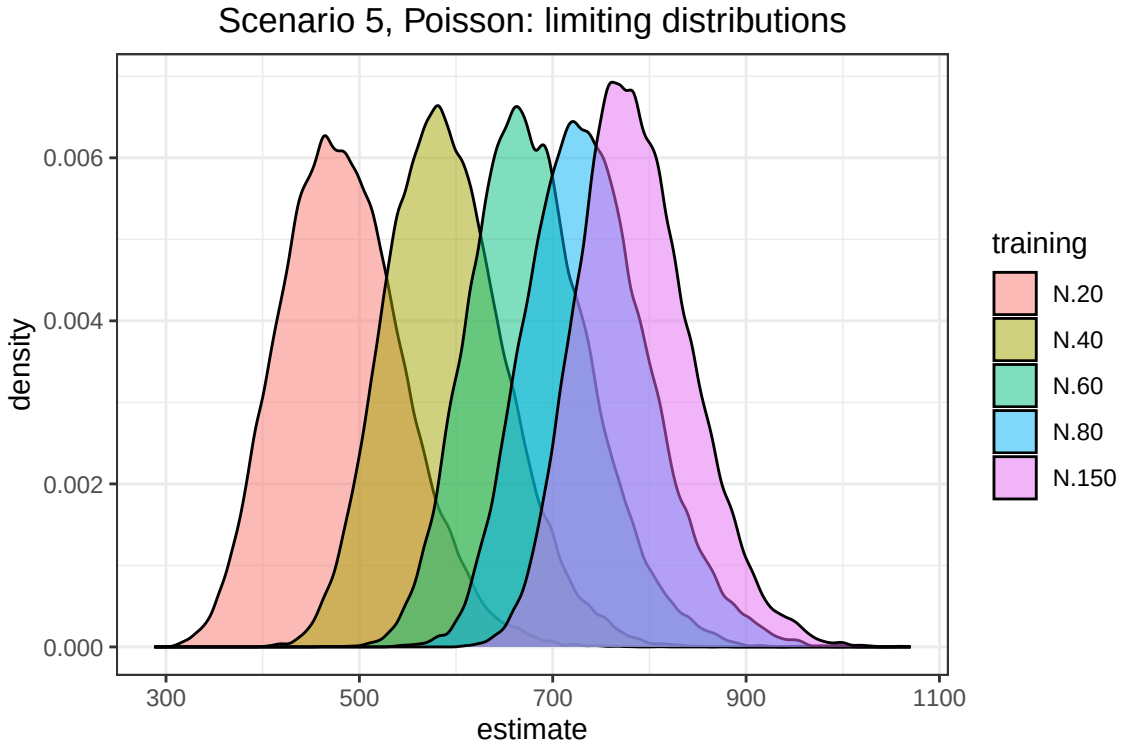
Here, we report (i.b) the extrapolated rarefaction curve (on a single dataset).

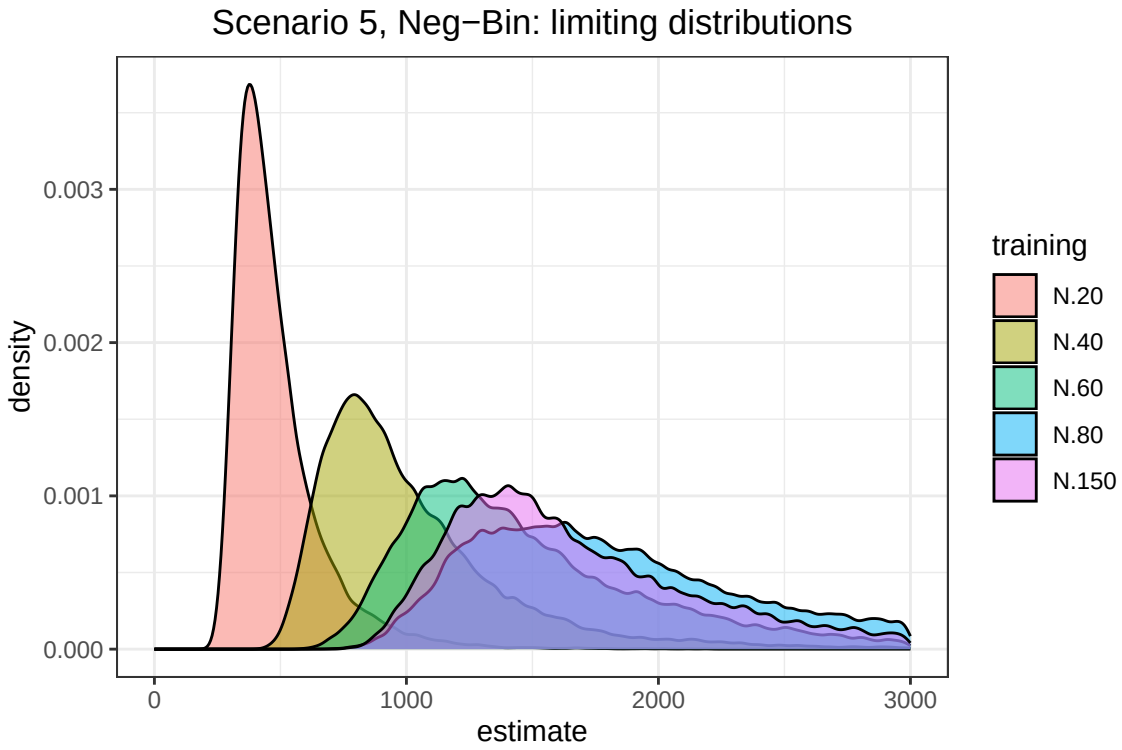


Here, we report (ii) the accuracy of the estimated number of unseen features in the test sample, over $D = 50$ replicated datasets.

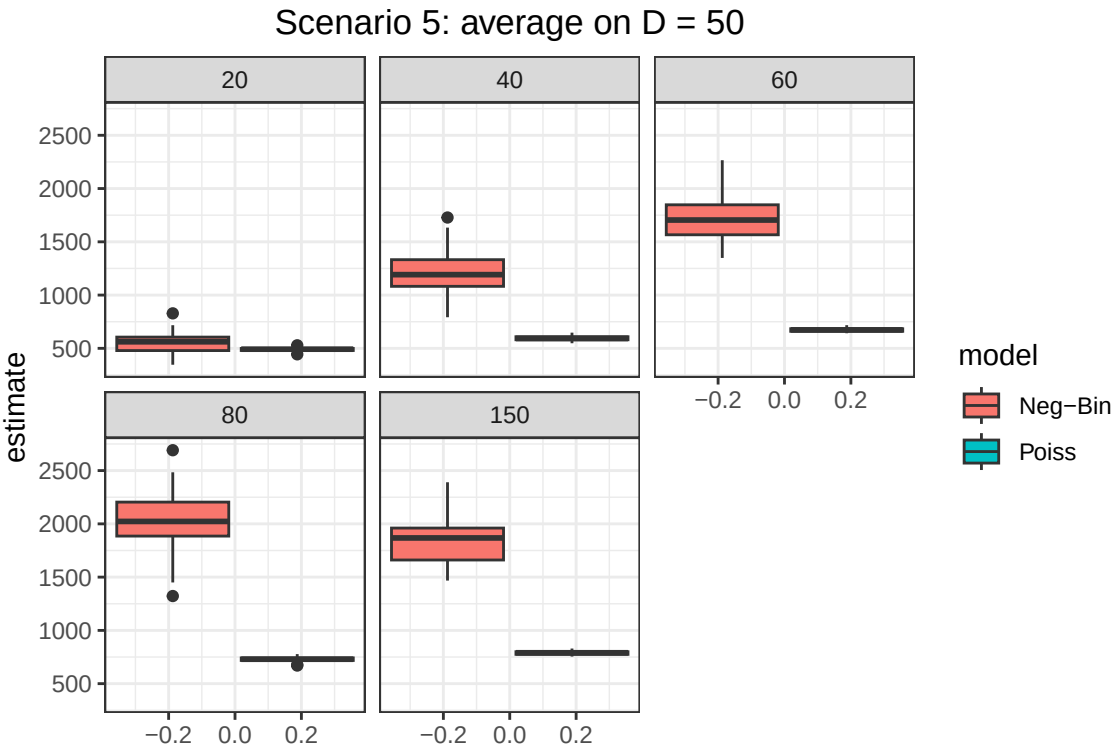


For the mixtures of Beta-Bernoulli, we report (iii) the posterior distribution of the total number of features (on a single dataset).





Finally, we report (iv) the expected value of the posterior distribution of the total number of features, over $D = 50$ replicated datasets.



Unbounded-features scenarios

We consider 3 different scenarios corresponding to 3 different growth rates of the number of distinct features observed in the population.

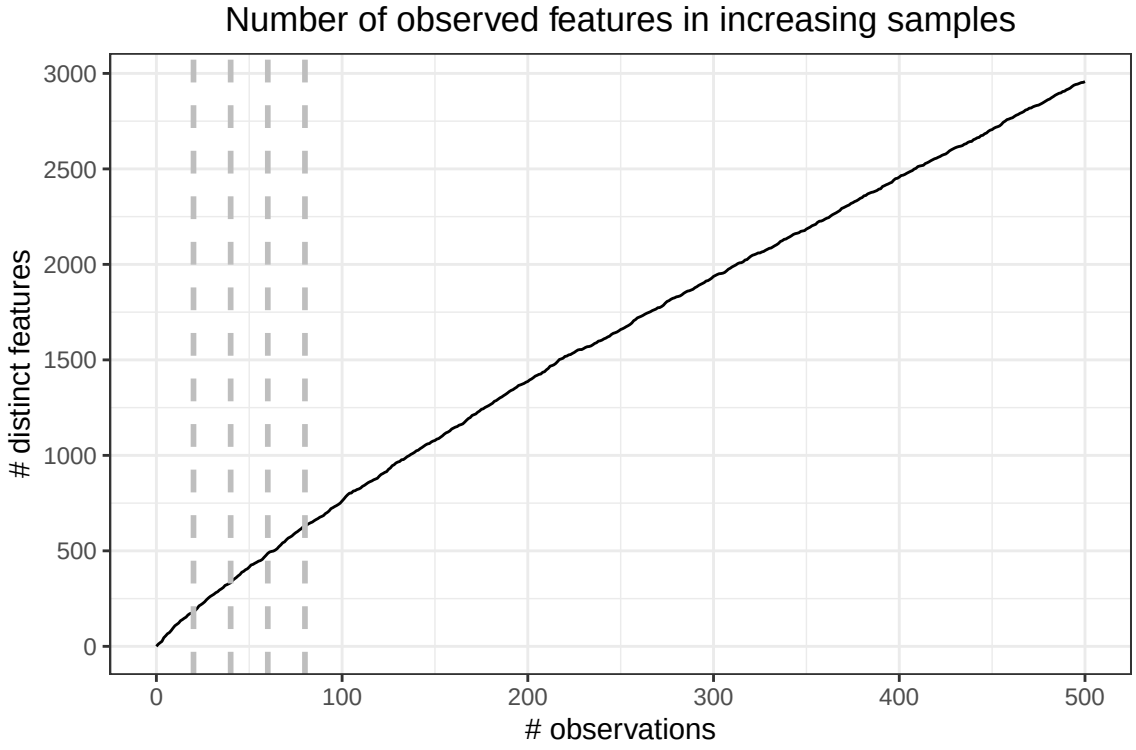
Scenario 1: polynomial growth with exponent 1

Set $\pi_k = \frac{1}{k}$, for $k = 1, \ldots, H$, with $H = 10^5$. Let L be the total dimension of the dataset, and consider different dimensions for the training set n, i.e.

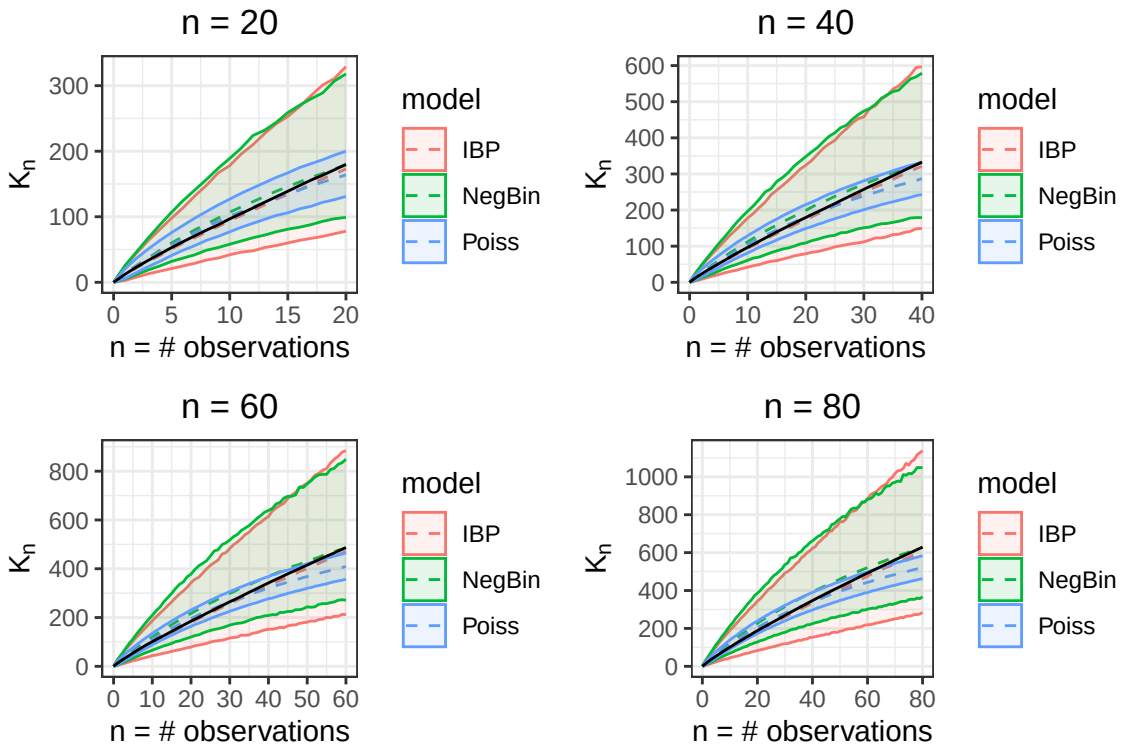
```
## L =  500
```

```
## n =  20  40  60  80
```

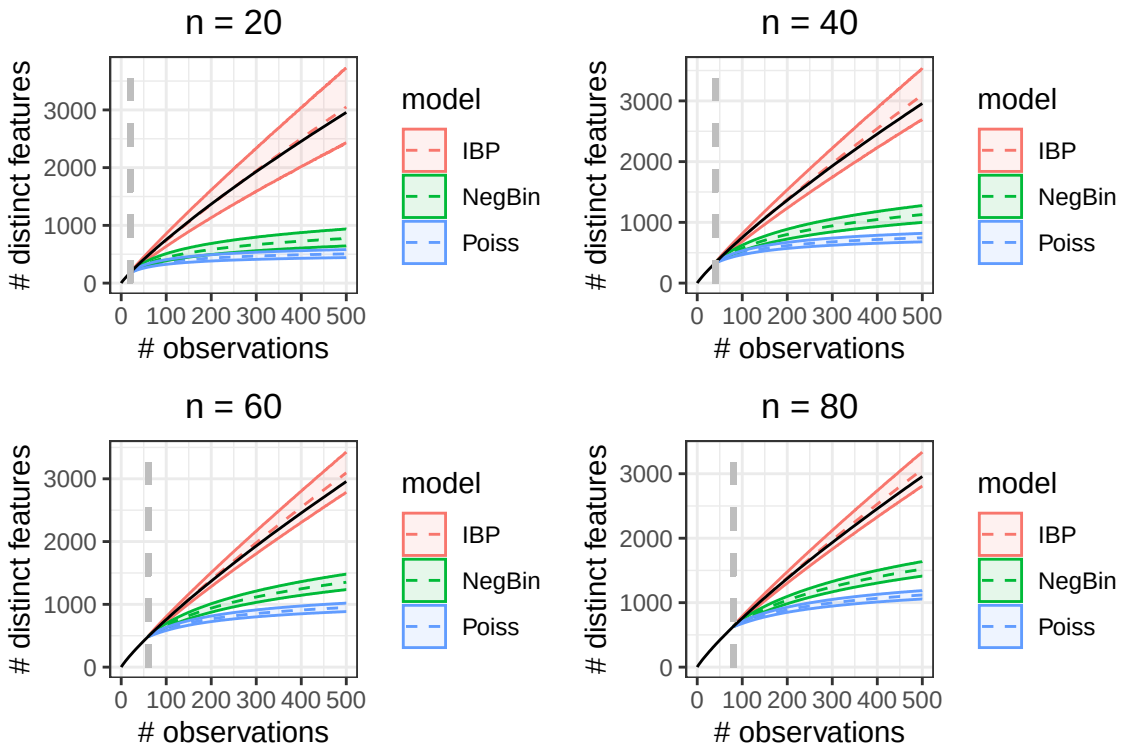
Here, the curve representing the number of observed features in increasing samples, where the grey vertical lines indicate the different training dimensions.



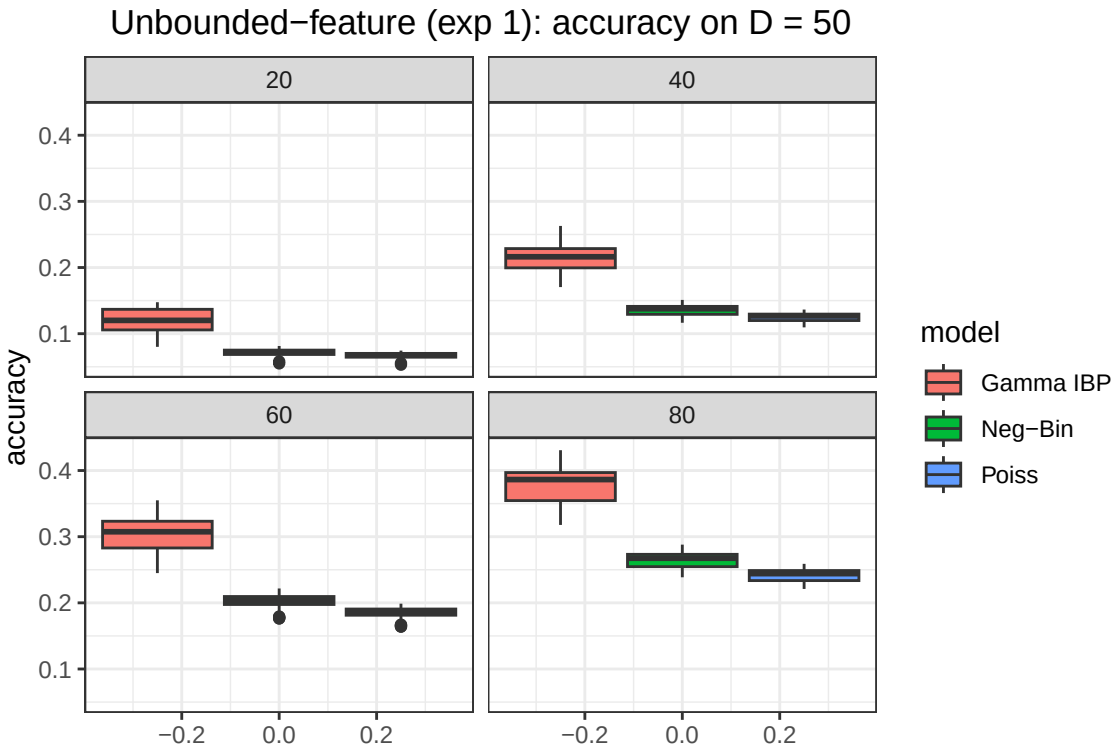
Here, we report (i.a) the in-sample rarefaction curve (on a single dataset).



Here, we report (i.b) the extrapolated rarefaction curve (on a single dataset).



Here, we report (ii) the accuracy of the estimated number of unseen features in the test sample, over $D = 50$ replicated datasets.



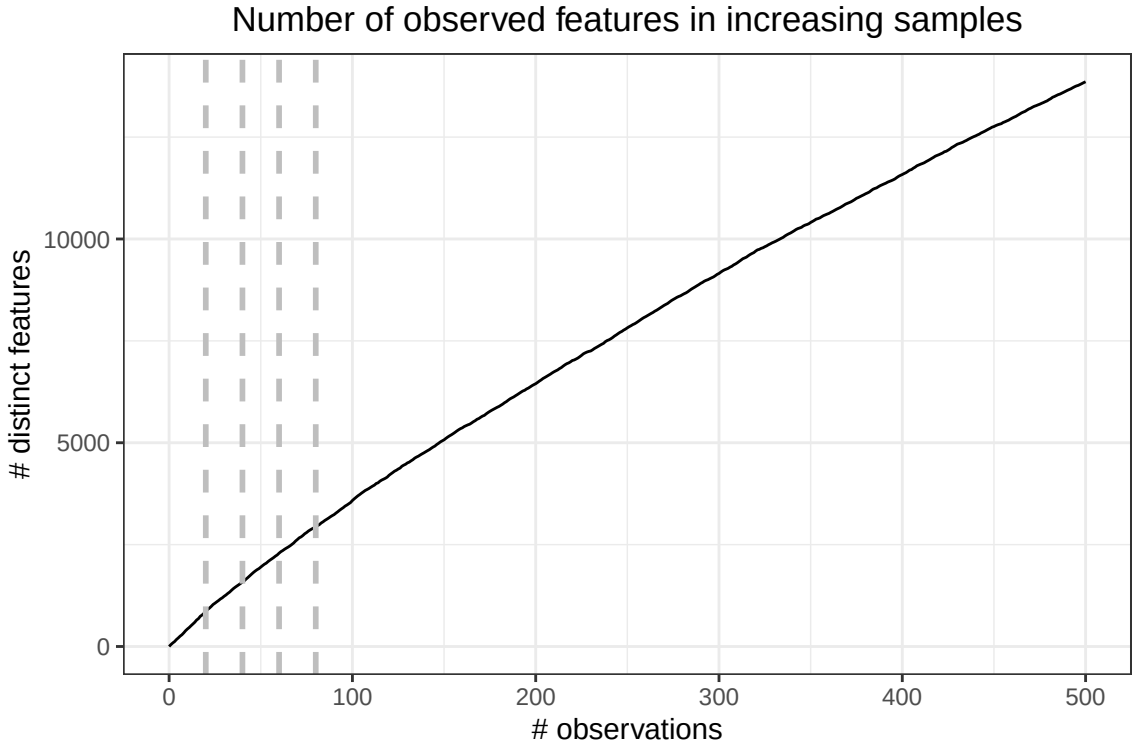
Scenario 2: polynomial growth with exponent 0.8

Set $\pi_k = \frac{1}{k}$, for $k = 1, \ldots, H$, with $H = 10^5$. Let L be the total dimension of the dataset, and consider different dimensions for the training set n, i.e.

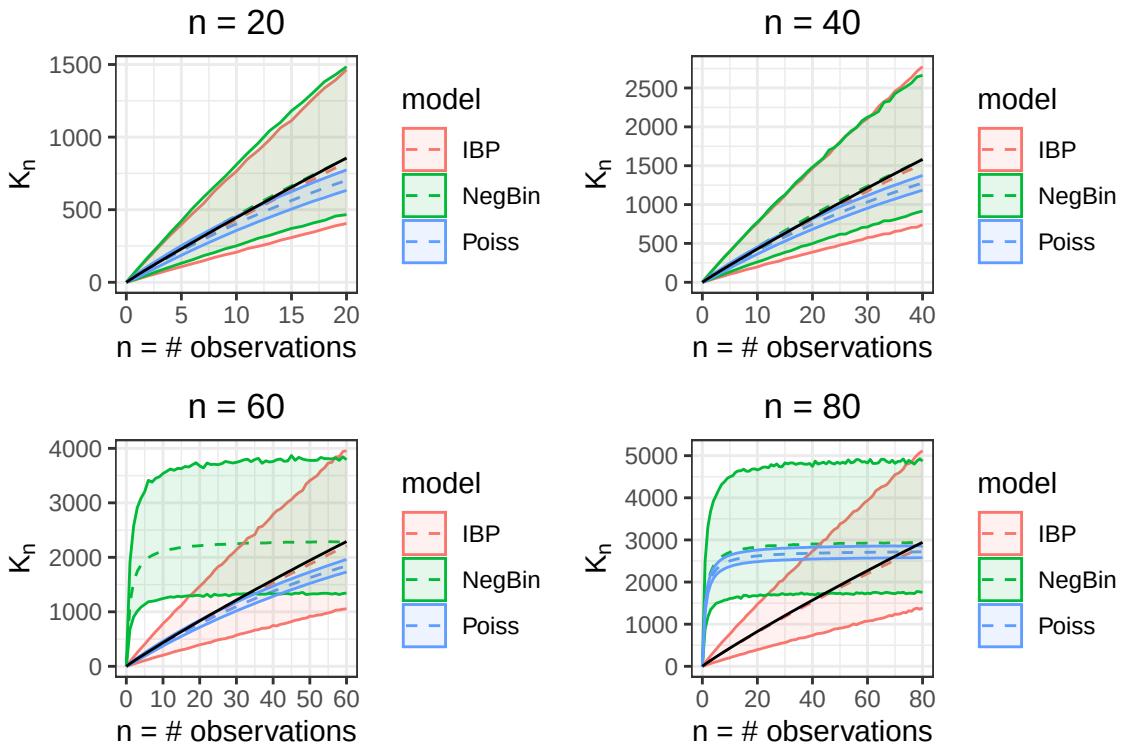
```
## L =  500
```

```
## n =  20  40  60  80
```

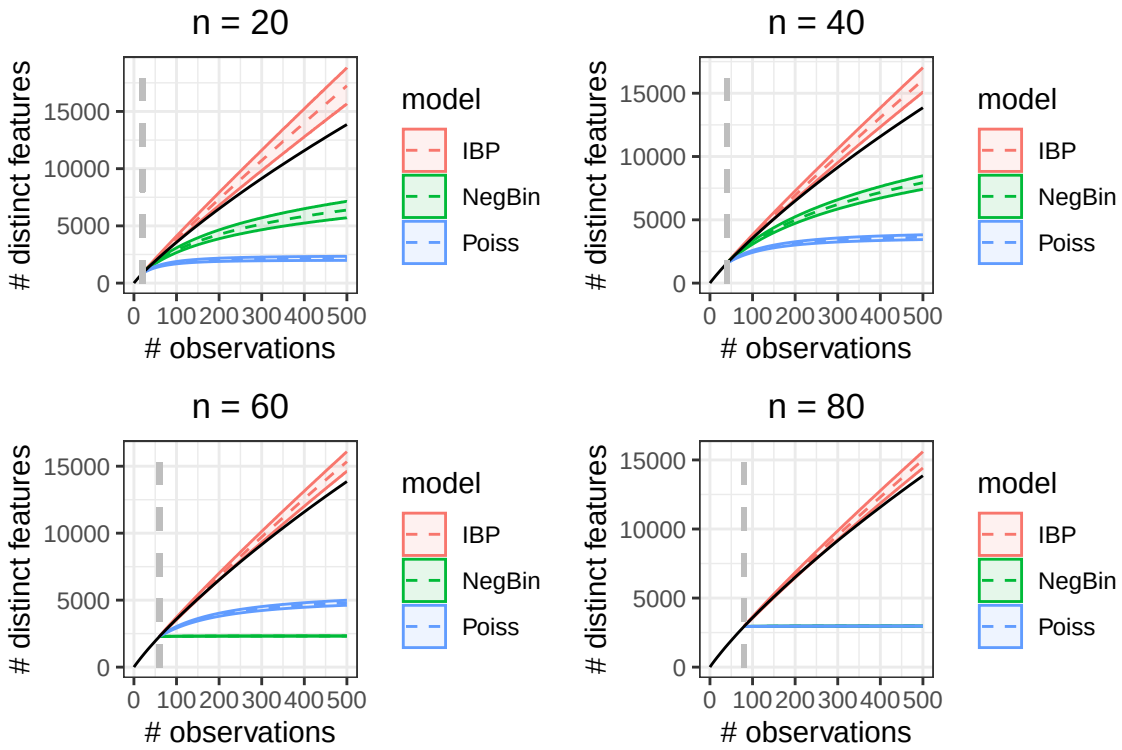
Here, the curve representing the number of observed features in increasing samples, where the grey vertical lines indicate the different training dimensions.



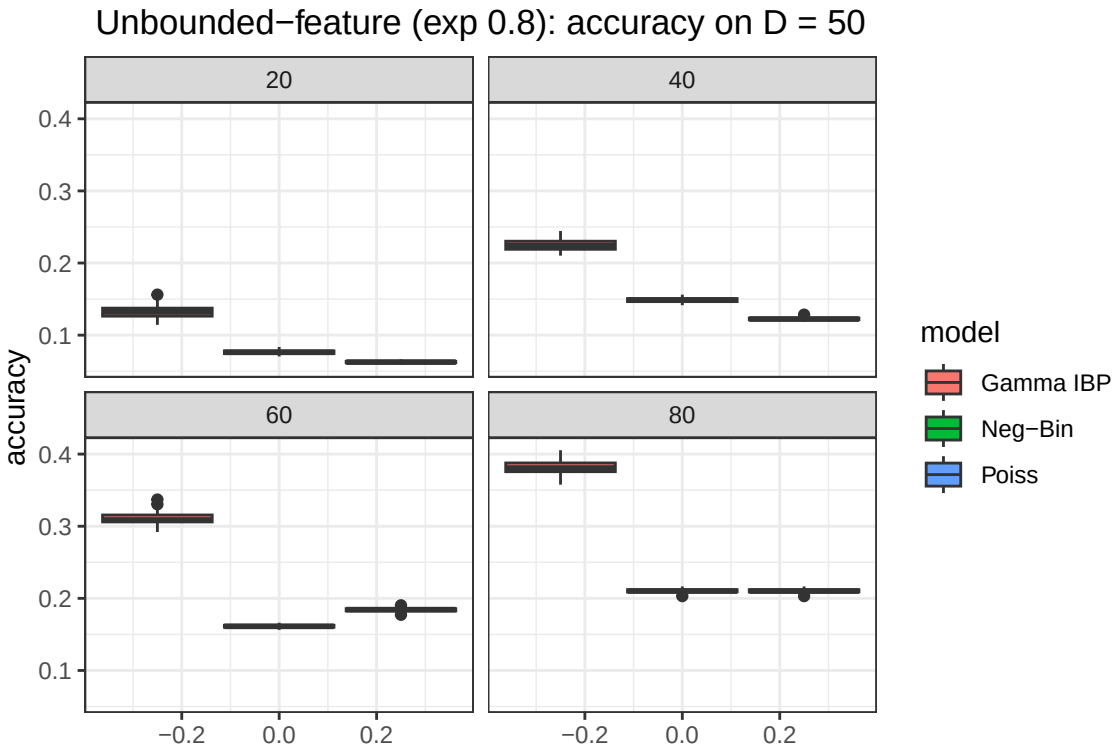
Here, we report (i.a) the in-sample rarefaction curve (on a single dataset).



Here, we report (i.b) the extrapolated rarefaction curve (on a single dataset).



Here, we report (ii) the accuracy of the estimated number of unseen features in the test sample, over $D = 50$ replicated datasets.



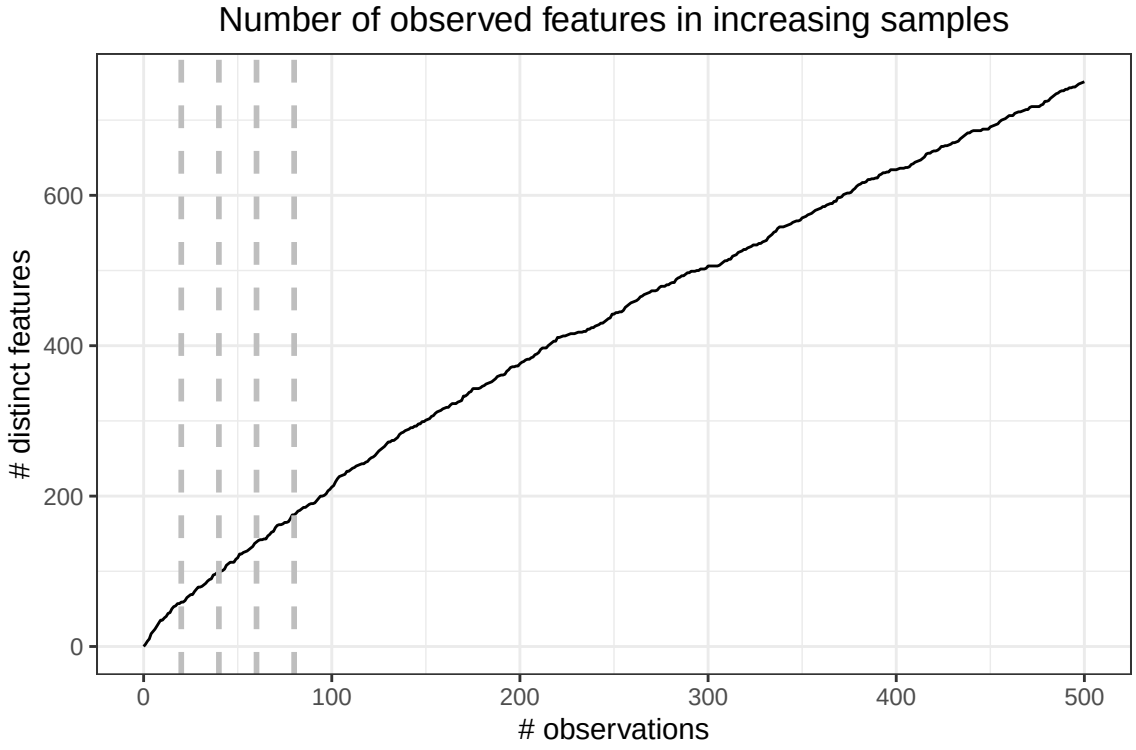
Scenario 3: polynomial growth with exponent 1.2

Set $\pi_k = \frac{1}{k}$, for $k = 1, \ldots, H$, with $H = 10^5$. Let L be the total dimension of the dataset, and consider different dimensions for the training set n, i.e.

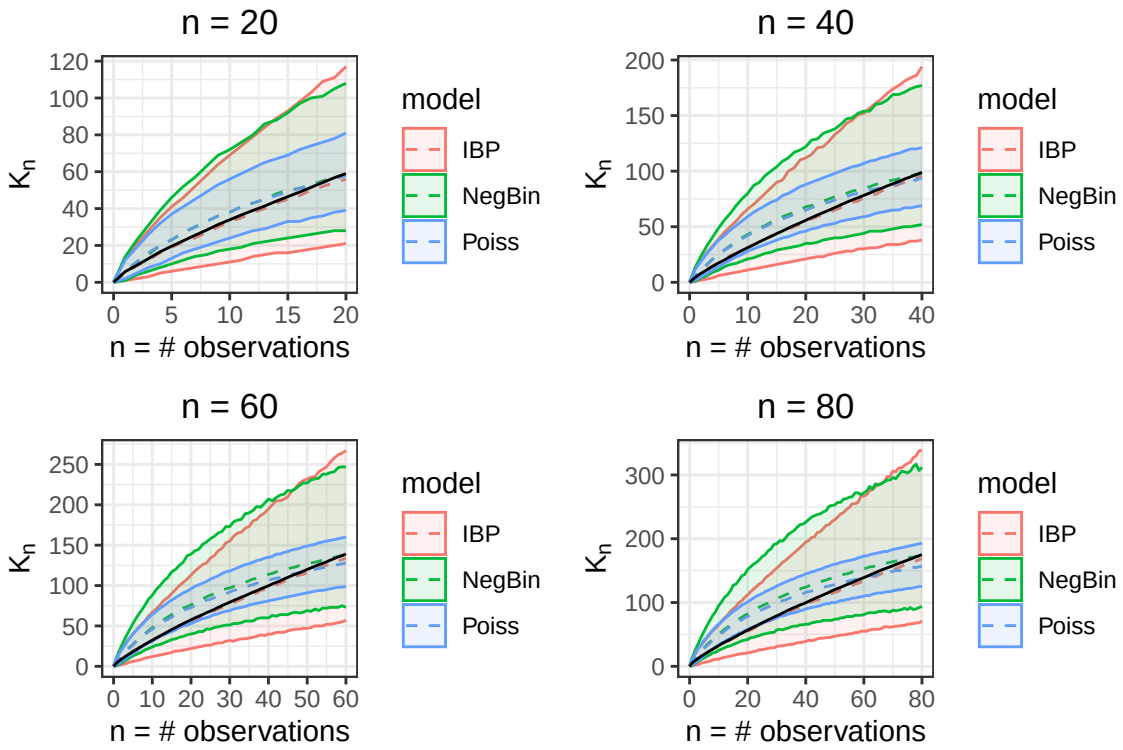
```
## L =  500
```

```
## n =  20  40  60  80
```

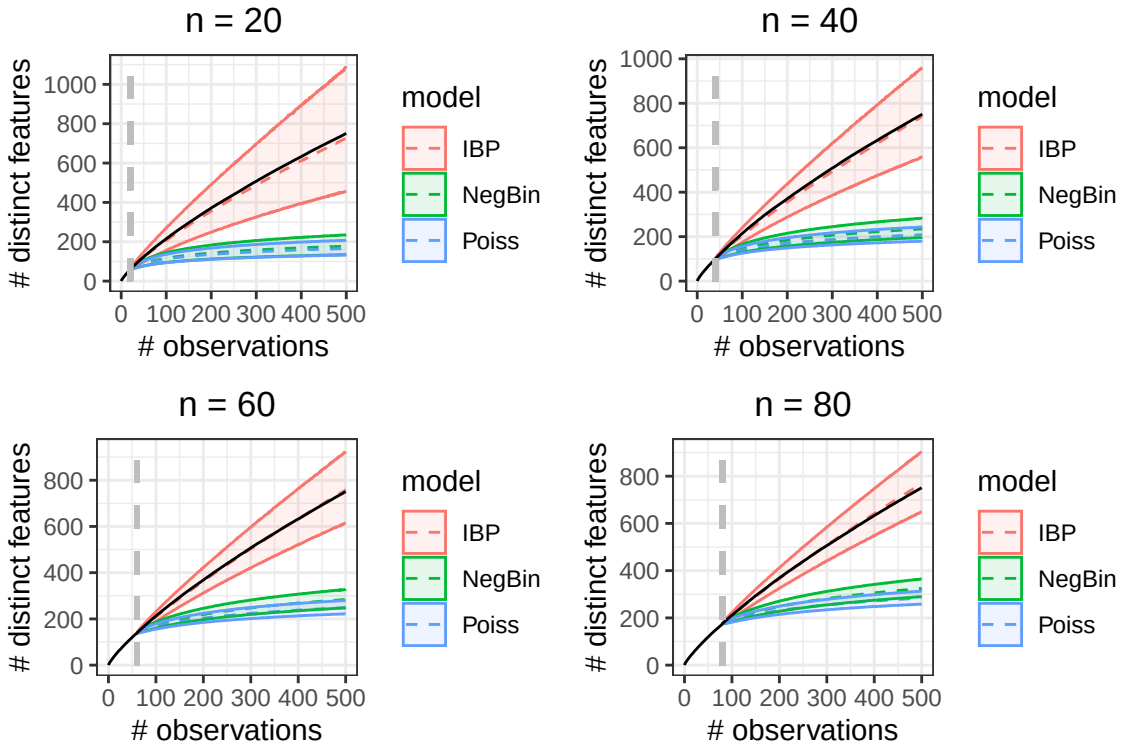
Here, the curve representing the number of observed features in increasing samples, where the grey vertical lines indicate the different training dimensions.



Here, we report (i.a) the in-sample rarefaction curve (on a single dataset).



Here, we report (i.b) the extrapolated rarefaction curve (on a single dataset).



Here, we report (ii) the accuracy of the estimated number of unseen features in the test sample, over $D = 50$ replicated datasets.