Approximate Bayesian Computation for Planet Occurrence Rate

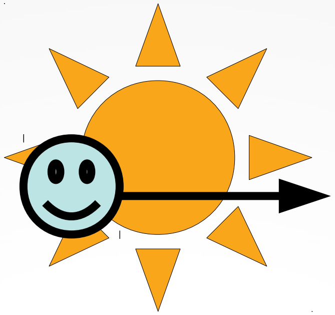
Hsu, D. C. et al. , 2018, The Astronomical Journal, 155:205.

Statistical Journal Club
14th November 2023

Makiko Ban

Introduction

Kepler transit mission



x 2700 exoplanets

Does *Kepler* data directly indicate the exoplanet population? NO!

- Orbital period
- Eccentricity
- Host-planet size ratio
- Signal to noise
- Target selection

Biases on the transit probability

Discovered exoplanets via transit by *Kepler*

Approximate Bayesian Computation

True exoplanet population

ABC method

First idea from Rubin, D. B. (1984)
Named and established by Beaumont, M. A. et al. (2002)

Standard Bayesian :

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

→ $$Posterior \propto Likelihood \times Prior$$

How to approximate likelihood function?

1) Simulate an artificial data with the arbitrary model parameters.
2) If 1) is close to the observed data, it is an acceptable model.
3) Accumulate acceptable models by repeating 1) – 2).
4) The final distribution of 3) indicates the model likelihoods.

Start ABC method gif (Leyshon, 2021)

ABC method

Key parameters for the ABC method:

- Summary statistics of the observed data ($S(Y_{obs})$)
- Summary statistics of the artificial data ($S(Y)$)
- Array of model parameters (ϕ)
- Distance function (ρ)
- Distance threshold (ϵ)

$$\pi_{ABC}(\phi|S) = \pi(\phi|\rho(S(Y), S(Y_{obs})) < \epsilon)$$

Application of Monte-Carlo simulation and Importance sampling:

- Sampling model parameter sets with a Monte-Carlo simulation provides $S(Y)$.
- ϵ can be updated by weighting the distribution of a model parameter from a previous ABC generation (Importance sampling).

ABC method for *Kepler* data

Goal

To find planet occurrence rates over different exoplanet periods and radii.

Period = from 0.5 to 320 days
Radius = from 0.5 to 16 R_{Earth}

Each bin = (period, radius)
is independent.

Procedure

1) Create a planetary catalogue with a given occurrence rate.
2) Simulate transit detection using the planetary catalogue from 1) and create an observed catalogue.
3) Derive summary statistics and distance function for each bin.
4) Gather a set of acceptable occurrence rates.
5) Find ABC posterior and update prior distribution and distance threshold for each bin.
6) Repeat 1) – 5) until it meets the ending condition.

ABC method for *Kepler* data

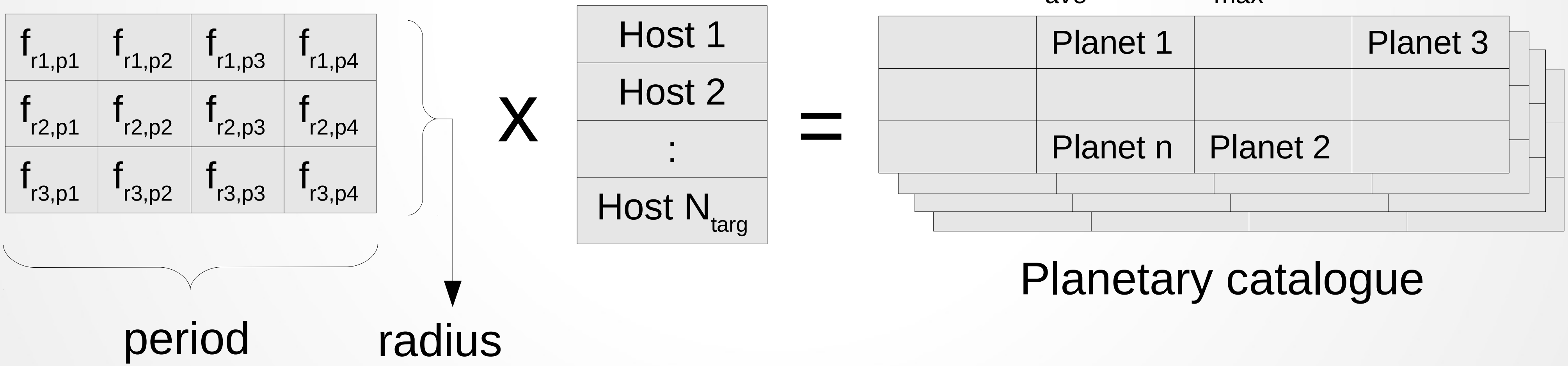
1) Create a planetary catalogue with a given occurrence rate.

$p_{r,p}$ = Prior distribution of occurrence rate per bin.

$f_{r,p}$ = Selected occurrence rate from $P_{r,p}$ per bin.

N_{targ} = 150,518 target host stars from the *Kepler* data.

$n_{s,r,p}$ = Number of planets per host star ← Poison($k \leq 10$, $\lambda = \sum_{r,p} f_{r,p}$).



$n_{ave} = \lambda$, $n_{max} = 10$

$f_{r1,p1}$	$f_{r1,p2}$	$f_{r1,p3}$	$f_{r1,p4}$
$f_{r2,p1}$	$f_{r2,p2}$	$f_{r2,p3}$	$f_{r2,p4}$
$f_{r3,p1}$	$f_{r3,p2}$	$f_{r3,p3}$	$f_{r3,p4}$

X

Host 1
Host 2
:
Host N_{targ}

=

	Planet 1		Planet 3
	Planet n	Planet 2	

Planetary catalogue

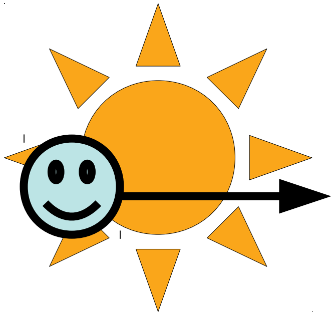
$$p_{r,p} = \begin{cases} \text{Inverse Detection Efficiency Method (IDEM)} & \text{for } 0^{th} \text{ generation} \\ \text{Importance Sampling distribution} & \text{for } >0^{th} \text{ generation} \end{cases}$$

ABC method for *Kepler* data

2) Simulate transit detection using the planetary catalogue from 1) and create an observed catalogue.

P = Period $^{(c)}$
d = Transit depth $^{(c)}$
D = Transit duration $^{(c)+\{i,e,\omega\}}$
t_0 = Time of 0th transit $^{(t)}$
i = Inclination $^{(p)}$
e = Eccentricity $^{(p)}$
ω = Angular velocity $^{(p)}$
b = *Transit impact parameter* $^{(t)}$

(c) value from the catalogue in 1).
(p) arbitrary value per planet
(t) arbitrary value per transit simulation



Detectable by *Kepler*?

YES

NO

Find “measured”

$$\hat{P}, \underbrace{\hat{d}, \hat{D}, \hat{t}_0}_{\hat{R}_p}$$

Relocate planets by $\{\hat{P}, \hat{R}_p\}$

Planet 2 :	: :	: :	: :
: :	Planet 4 :	Planet 1 :	: Planet n
: :	Planet 3 :	: :	: :

Observed catalogue

ABC method for *Kepler* data

3) Derive summary statistics and distance function for each bin.

N_{targ} = 150,518 target host stars from the *Kepler* data.
N_p = 3380 planet candidates for N_{targ} from the *Kepler* data.
$n_{r,p}$ = Number of planets per bin in the observed catalogue.

Planet 2 :	: :	: :	: :
: :	Planet 4 :	Planet 1 :	: Planet n
: :	Planet 3 :	: :	: :

Observed catalogue

$n_{r1,p1}$	$n_{r1,p2}$	$n_{r1,p3}$	$n_{r1,p4}$
$n_{r2,p1}$	$n_{r2,p2}$	$n_{r2,p3}$	$n_{r2,p4}$
$n_{r3,p1}$	$n_{r3,p2}$	$n_{r3,p3}$	$n_{r3,p4}$

Number of observed planets

Summary statistics of a model :

$$S(Y)_{r,p}=\frac{n_{r,p}}{N_{targ}}$$

Summary statistics of *Kepler* data :

$$S(Y_{obs})_{r,p}=\frac{\{N_{(r,p)}\subset N_p\}}{N_{targ}}$$

Distance function :

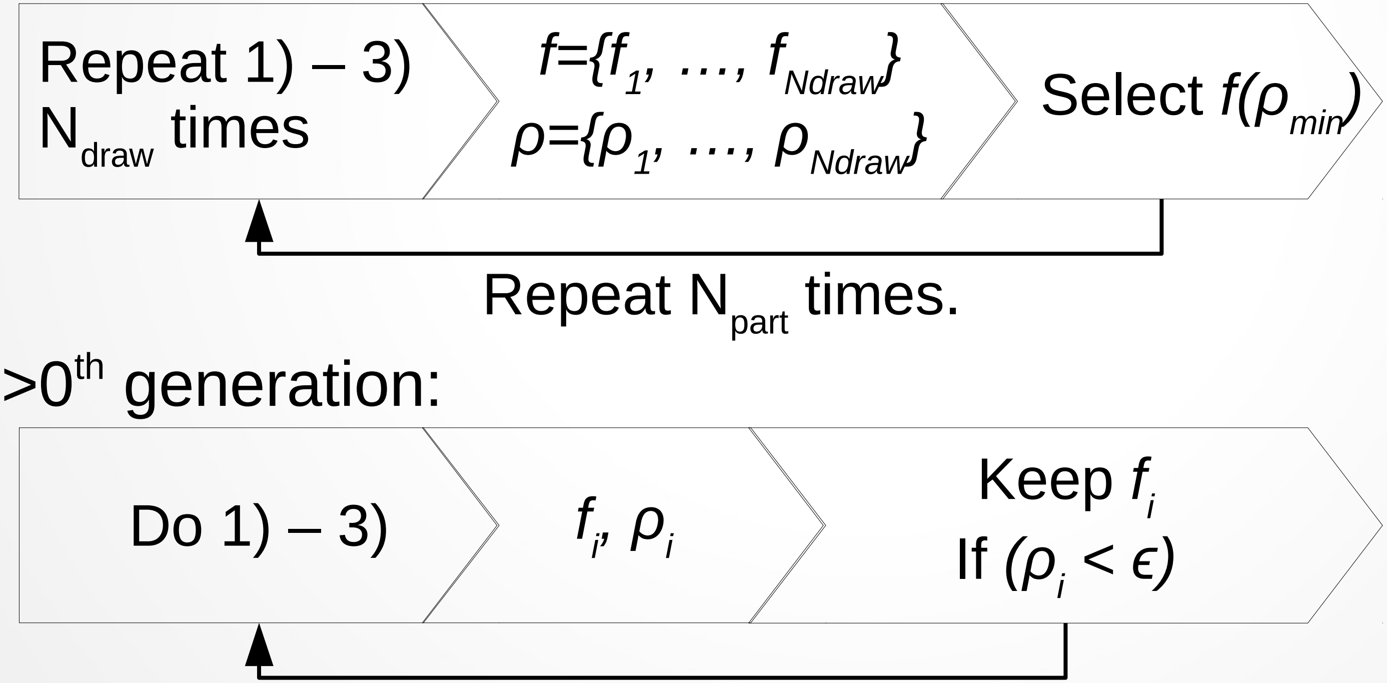
$$\rho(S(Y_{obs}),S(Y))_{r,p}=\{S(Y)-S(Y)\}^2$$

ABC method for *Kepler* data

4) Gather a set of acceptable occurrence rates.

N_{draw}, N_{part} = Arbitrary number of trials and survivals.
f_i = A model occurrence rate (*i*).
ρ_i = A distance function of a model occurrence rate (*i*).
ϵ = Distance threshold

At 0th generation:



Repeat 1) – 3) N_{draw} times

$f=\{f_1, \dots, f_{Ndraw}\}$
$\rho=\{\rho_1, \dots, \rho_{Ndraw}\}$

Select $f(\rho_{min})$

Repeat N_{part} times.

At >0th generation:

Do 1) – 3)

f_i, ρ_i

Keep f_i
If $(\rho_i < \epsilon)$

Repeat until N_{part} models are gained. ← $N_{draw} \leq N_{max}$ trials.

Acceptable Occurrence rate

f_1, ρ_1
f_2, ρ_2
:
f_{Npart}, ρ_{Npart}

In a bin

ABC method for *Kepler* data

5) Find ABC posterior and update prior distribution and distance threshold for each bin.

f_i, ρ_i = An occurrence rate and its distance function.
σ = Standard deviation of acceptable occurrence rates.
$N_f(f_i,\sigma)$ = Probability function of an occurrence rate in a form of normal distribution.
w_i = Normalised sampling weight of an occurrence rate.
$p_{0,r,p}$ = Prior distribution function at 0th generation (IDEM).
$p_{r,p}$ = Prior distribution function at current generation.

Sampling weight : $\boxed{w_i^* = \frac{p_{0,r,p}(f_i,\sigma)}{p_{r,p}(f_i,\sigma)}}$ → Normalised : $w_i = \frac{w_i^*}{\sum_i^{N_{part}} w_i^*}$

ABC posterior : $\boxed{p_{ABC}(f|Y_{obs})_{r,p} = \sum_i^{N_{part}} w_i N_f(f_i,\sigma)_{r,p}}$

Prior distribution for next generation : $\boxed{p_{(IS)r,p} = \sum_i^{N_{part}} w_i N_f(f_i,\tau\sigma)_{r,p}}$ where $\tau \simeq 2$

Distance threshold for next generation : $\boxed{\epsilon = max(\rho_i)}$

ABC method for *Kepler* data

6) Repeat 1) – 5) until it meets the ending condition.

Ending conditions :

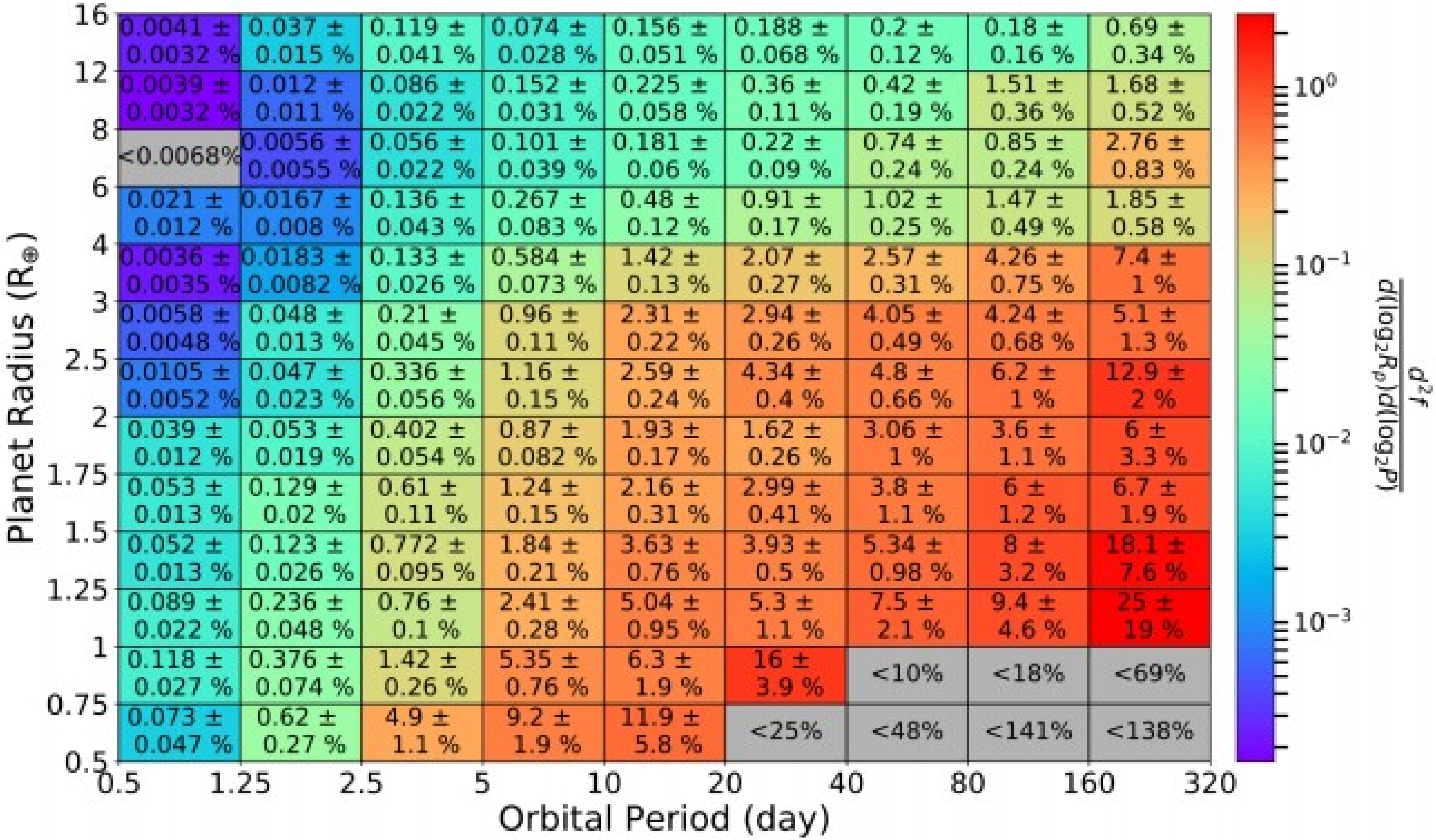
(1) Mean distance function is less than the distance threshold.
(2) The ABC generation reaches to 200.
(3) Number of same occurrence rate drawing > N_{part}.
(4) Median number of N_{draw} > $0.2*N_{max}$.
(5) No improvement of ϵ in consecutive three generations.

The finalised ABC posterior tells the likelihood of the occurrence rate!

Uncertainty of the ABC posterior :

- Reaches to Monte-Carlo noise limit.
- Shorter period bin yields smaller uncertainty.

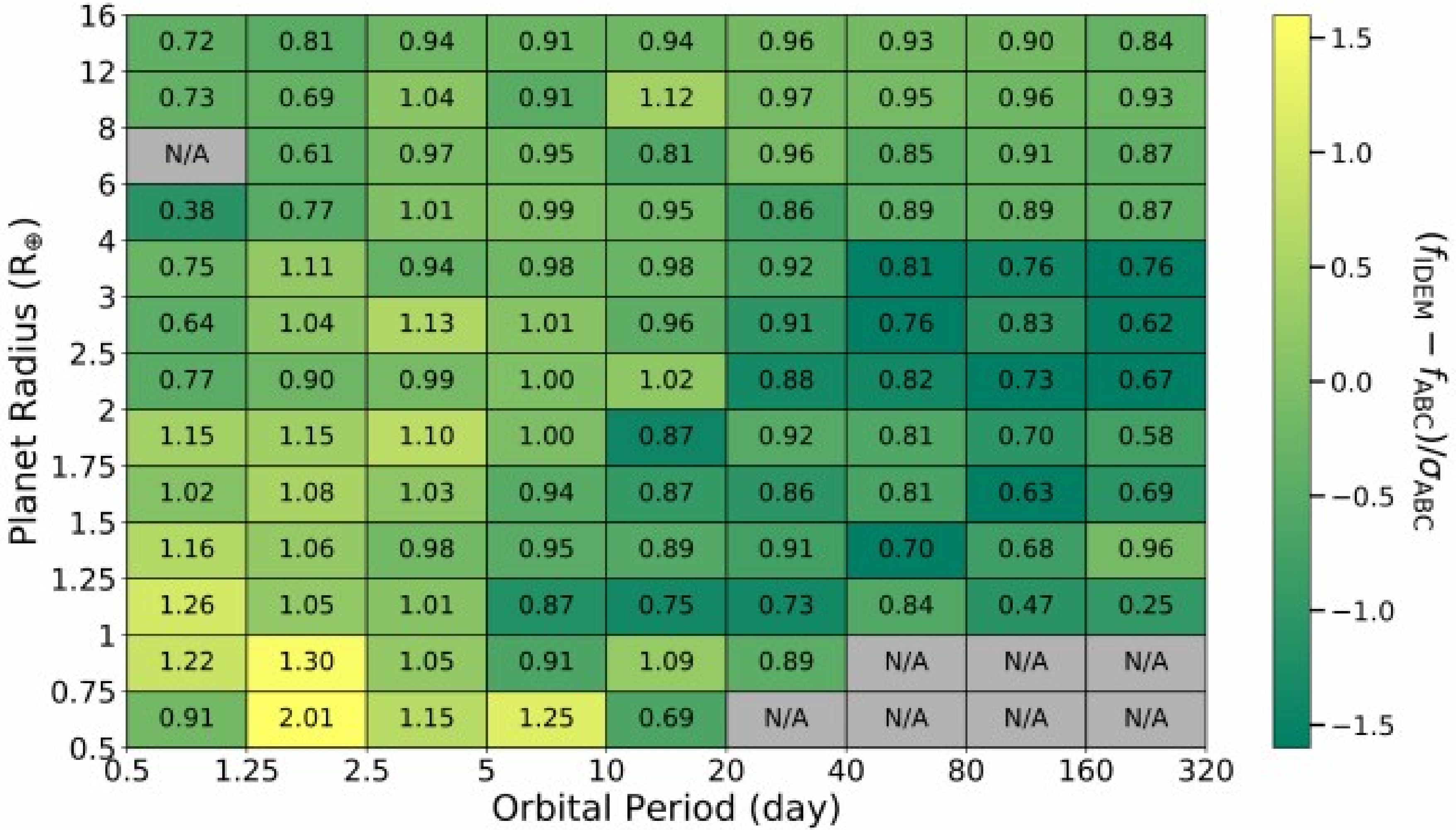
ABC method for *Kepler* data



0.0041 ± 0.0032 %	0.037 ± 0.015 %	0.119 ± 0.041 %	0.074 ± 0.028 %	0.156 ± 0.051 %	0.188 ± 0.068 %	0.2 ± 0.12 %	0.18 ± 0.16 %	0.69 ± 0.34 %
0.0039 ± 0.0032 %	0.012 ± 0.011 %	0.086 ± 0.022 %	0.152 ± 0.031 %	0.225 ± 0.058 %	0.36 ± 0.11 %	0.42 ± 0.19 %	1.51 ± 0.36 %	1.68 ± 0.52 %
<0.0068%	0.0056 ± 0.0055 %	0.056 ± 0.022 %	0.101 ± 0.039 %	0.181 ± 0.06 %	0.22 ± 0.09 %	0.74 ± 0.24 %	0.85 ± 0.24 %	2.76 ± 0.83 %
0.021 ± 0.012 %	0.0167 ± 0.008 %	0.136 ± 0.043 %	0.267 ± 0.083 %	0.48 ± 0.12 %	0.91 ± 0.17 %	1.02 ± 0.25 %	1.47 ± 0.49 %	1.85 ± 0.58 %
0.0036 ± 0.0035 %	0.0183 ± 0.0082 %	0.133 ± 0.026 %	0.584 ± 0.073 %	1.42 ± 0.13 %	2.07 ± 0.27 %	2.57 ± 0.31 %	4.26 ± 0.75 %	7.4 ± 1 %
0.0058 ± 0.0048 %	0.048 ± 0.013 %	0.21 ± 0.045 %	0.96 ± 0.11 %	2.31 ± 0.22 %	2.94 ± 0.26 %	4.05 ± 0.49 %	4.24 ± 0.68 %	5.1 ± 1.3 %
0.0105 ± 0.0052 %	0.047 ± 0.023 %	0.336 ± 0.056 %	1.16 ± 0.15 %	2.59 ± 0.24 %	4.34 ± 0.4 %	4.8 ± 0.66 %	6.2 ± 1 %	12.9 ± 2 %
0.039 ± 0.012 %	0.053 ± 0.019 %	0.402 ± 0.054 %	0.87 ± 0.082 %	1.93 ± 0.17 %	1.62 ± 0.26 %	3.06 ± 1 %	3.6 ± 1.1 %	6 ± 3.3 %
0.053 ± 0.013 %	0.129 ± 0.02 %	0.61 ± 0.11 %	1.24 ± 0.15 %	2.16 ± 0.31 %	2.99 ± 0.41 %	3.8 ± 1.1 %	6 ± 1.2 %	6.7 ± 1.9 %
0.052 ± 0.013 %	0.123 ± 0.026 %	0.772 ± 0.095 %	1.84 ± 0.21 %	3.63 ± 0.76 %	3.93 ± 0.5 %	5.34 ± 0.98 %	8 ± 3.2 %	18.1 ± 7.6 %
0.089 ± 0.022 %	0.236 ± 0.048 %	0.76 ± 0.1 %	2.41 ± 0.28 %	5.04 ± 0.95 %	5.3 ± 1.1 %	7.5 ± 2.1 %	9.4 ± 4.6 %	25 ± 19 %
0.118 ± 0.027 %	0.376 ± 0.074 %	1.42 ± 0.26 %	5.35 ± 0.76 %	6.3 ± 1.9 %	16 ± 3.9 %	<10%	<18%	<69%
0.073 ± 0.047 %	0.62 ± 0.27 %	4.9 ± 1.1 %	9.2 ± 1.9 %	11.9 ± 5.8 %	<25%	<48%	<141%	<138%

Figure 5. ABC estimated occurrence rates for the Q1–Q16 planet candidates orbiting FGK stars using the Christiansen et al. (2015) gamma CDF curve for the detection efficiency. The numerical values of the occurrence rates are stated as percentages (i.e., 10^{-2}). The color coding of each cell is based on $(d^2f)/[d(\log_2 R_p)\, d(\log_2 P)]$, which provides an occurrence rate normalized to the width of the bin and therefore is not dependent on choice of grid density. Cells colored gray have estimated upper limits for the occurrence rate. Note that the bin sizes are not constant.

ABC method for *Kepler* data



0.72	0.81	0.94	0.91	0.94	0.96	0.93	0.90	0.84
0.73	0.69	1.04	0.91	1.12	0.97	0.95	0.96	0.93
N/A	0.61	0.97	0.95	0.81	0.96	0.85	0.91	0.87
0.38	0.77	1.01	0.99	0.95	0.86	0.89	0.89	0.87
0.75	1.11	0.94	0.98	0.98	0.92	0.81	0.76	0.76
0.64	1.04	1.13	1.01	0.96	0.91	0.76	0.83	0.62
0.77	0.90	0.99	1.00	1.02	0.88	0.82	0.73	0.67
1.15	1.15	1.10	1.00	0.87	0.92	0.81	0.70	0.58
1.02	1.08	1.03	0.94	0.87	0.86	0.81	0.63	0.69
1.16	1.06	0.98	0.95	0.89	0.91	0.70	0.68	0.96
1.26	1.05	1.01	0.87	0.75	0.73	0.84	0.47	0.25
1.22	1.30	1.05	0.91	1.09	0.89	N/A	N/A	N/A
0.91	2.01	1.15	1.25	0.69	N/A	N/A	N/A	N/A

Figure 6. Ratios of the planet occurrence rate inferred by the IDEM to the planet occurrence rate inferred by the ABC method based on the Q1–Q16 candidates orbiting FGK stars. The color indicates the relative difference between the occurrence rates scaled by the uncertainty given by the standard deviation from the ABC method. The two methods give similar results for easily detected planets (most bins with high S/N are within ~1 standard deviations), but the IDEM substantially underestimates the frequency of near-Earth-size planets for orbital periods beyond ~80 days.

* f_{IDEM} : Same as the means of the prior distribution at 0^{th} generation.

Summary

Approximate Bayesian Computation

- Monte-Carlo method using a distribution of models.
- The likelihood is approximated by accumulating “good” models.
- A final posterior indicates the likelihood distribution of the models.

Exoplanet occurrence rate from *Kepler* data with ABC application

- The uncertainty of the ABC posterior is as small as the Monte-Carlo noise limit.
- The occurrence rates for Earth-like planets > 80 days period is larger than the former expectation.

Reference

- Beaumont, M.A., et al., 2002, *Genetics*, 162(4): 2025-2035.
- Christiansen, J. L., et al., 2015, *ApJ*, 810(2): 95.
- Hsu, D. C., et al., 2018, *Astronomical Joutnal*, 155:205.
- Leyshon, T., 2021, Towards Data science, https://towardsdatascience.com/the-abcs-of-approximate-bayesian-computation-bfe11b8ca341
- Rubin, D. B., 1984, *Ann. Statist*, 12(4): 1151-1172.