Draft version November 9, 2016
Preprint typeset using LaTeX style AASTeX6
With modifications by David W. Hogg and Daniel Foreman-Mackey

APPROXIMATE BAYESIAN COMPUTATION OF THE POPULATION OF EXOPLANETS

Daniel Foreman-Mackey[1,2], and friends

[1]danfm@uw.edu; Sagan Fellow
[2]Astronomy Department, University of Washington, Seattle, WA, 98195, USA

ABSTRACT

The *Kepler* Mission was designed to statistical questions about the population of exoplanets. So far, all occurrence rate and distribution results have required either strong modeling assumptions or heuristic (qualitative) inference methods. We present a general method of occurrence rate measurement using Approximate Bayesian Computation (ABC). ABC allows rigorous posterior inference of population parameters without computing a likelihood function. Instead, we only need to be able to simulate representative data sets. We demonstrate the importance of this method on a simple toy problem with a tractable likelihood function and then apply the method to a catalog of exoplanet discoveries from the *Kepler* Mission to simultaneously measure the occurrence rate and period, size, multiplicity, and mutual inclination distributions.

Keywords: methods: data analysis — methods: statistical — catalogs — planetary systems — stars: statistics

1. INTRODUCTION

- Exoplanet populations
- Why Approximate Bayesian Computation (ABC)? Which problems? Limitations of current methods.

2. A MOTIVATING EXAMPLE: INFERENCE WITH A POISSON PROCESS

A standard problem in the astronomy literature that can – and has been – solved using Approximate Bayesian Computation (ABC) is the situation where we have a procedure for simulating a population of objects that we then want to compare to the observed population. For example, in the exoplanet literature, this is the case for

most studies of the multiplicity distribution of exoplanets; it is easy to simulate and observe a population of exoplanetary systems with a given multiplicity distribution (for example Fang & Margot 2012; Ballard & Johnson 2016) *DFM*: CITE Jack but it is difficult to write down an analytic or otherwise tractable likelihood function without making strong assumptions (Tremaine & Dong 2012, for example). It is common practice in this field to use the simulator to do inference by comparing the simulations to the observed distributions using a heuristic "likelihood function". For example, Ballard & Johnson (2016) (*DFM*: Others?) draw a realization of the observed multiplicity distribution from the model and then compare this to the data by computing the Poisson likelihood of the observations but treating the simulated number as the expected rate. In other words, for each bin k in multiplicity, they compute a likelihood:

$$p(N_{\mathrm{obs},k} \mid \theta) = \frac{N_{\mathrm{sim},k}(\theta)^{N_{\mathrm{obs},k}} \, \exp\left[-N_{\mathrm{sim},k}(\theta)\right]}{N_{\mathrm{obs},k}!} \tag{1}$$

where $N_{\mathrm{sim},k} \sim p(N_{\mathrm{sim},k} \mid \theta)$ is the simulated number of systems with k transiting planets. This might seem intuitive but, as we will demonstrate shortly, it is not valid even if the underlying process is actually Poisson and using a model like this will lead to incorrect inferences.

To demonstrate the issues with this method, let's start with the simplest toy problem: a single observation of a Poisson process. In this case, the correct posterior inference can be performed because the correct likelihood is analytically tractable. The generative model is

$$N_{\mathrm{obs}} \sim p(N_{\mathrm{obs}} \mid \mu) = \frac{\mu^{N_{\mathrm{obs}}} \, e^{-\mu}}{N_{\mathrm{obs}}!} \tag{2}$$

with a single parameter μ. Placing a uniform prior on $\log \mu$ between μ_{min} and μ_{max}, the posterior probability becomes

$$p(\mu \mid N_{\mathrm{obs}}) \propto \begin{cases} \frac{1}{\mu} \frac{\mu^{N_{\mathrm{obs}}} \, e^{-\mu}}{N_{\mathrm{obs}}!} & \text{if } \mu_{\mathrm{min}} \le \mu \le \mu_{\mathrm{max}} \\ 0 & \text{otherwise} \end{cases} . \tag{3}$$

To test the methods, we generated a simulated "data set" of $N_{\mathrm{obs}} = 6$ from a Poisson process with $\mu_{\mathrm{true}} = 6.78$. Using this simulated data, Figure 1 compares the inferences made using Equation (1) to the exact posterior inference using Equation (3) and to inferences made using ABC with a pseudo-likelihood (the details will be given in the following section)

$$p(N_{\mathrm{obs}} \mid \mu) = \int K(N_{\mathrm{sim}}, \, N_{\mathrm{obs}}; \, \epsilon) \, p(N_{\mathrm{sim}} \mid \mu) \, \mathrm{d}N_{\mathrm{sim}} \tag{4}$$

where

$$p(N_{\mathrm{sim}} \mid \mu) = \frac{\mu^{N_{\mathrm{sim}}} \, e^{-\mu}}{N_{\mathrm{sim}}!} \tag{5}$$

is the generative model. For this demonstration, we use a Gaussian kernel

$$K(N_{\rm sim},\, N_{\rm obs};\, \epsilon) \propto \exp\left(-\frac{(N_{\rm sim} - N_{\rm obs})^2}{2\,\epsilon^2}\right) \tag{6}$$

with $\epsilon \equiv 10^{-3}$. The pseudo-likelihood is computed numerically as

$$p(N_{\rm obs} \mid \mu) \approx \frac{1}{S}\sum_{s=1}^{S} K({N_{\rm sim}}^{(s)},\, N_{\rm obs};\, \epsilon) \tag{7}$$

for ${N_{\rm sim}}^{(s)} \sim {\rm Pois}(\mu)$ and $S = 50$.

From Figure 1, it is clear that the ABC inference correctly recovers the exact posterior while the heuristic method underestimates the mean of the posterior and overestimates the variance. In practice, this inconsistency increases when the true underlying distribution is no longer Poisson.

The main point of this demonstration is that the heuristic method commonly used to do inference with stochastic population models is not sufficient if we want to do rigorous quantitative population inference. Instead, ABC can be used to capture the exact posterior.

3. A SHORT INTRODUCTION TO APPROXIMATE BAYESIAN COMPUTATION

The problem statement for ABC is that you have a method of (stochastically) simulating representative data from the model given a set of parameters but it is computationally intractable to compute the likelihood of a given data set conditioned on these same parameters. In other words, you can sample $D_{\rm sim} \sim p(D_{\rm sim} \mid \theta)$ but you can't directly evaluate $p(D_{\rm obs} \mid \theta)$ for any values $D_{\rm obs}$ and θ. This means that you can't use standard Markov Chain Monte Carlo (MCMC) methods to draw samples from the posterior probability $p(\theta \mid D_{\rm obs}) \propto p(D_{\rm obs} \mid \theta)\, p(\theta)$.

To motivate ABC, let's start with an extreme method. Let's say that sample a set of parameters $\theta \sim p(\theta)$ from your prior and then simulate some data $D_{\rm sim}$ that just happens to *exactly* match the observed data $D_{\rm obs}$. In this case, under your simulation model, this value of θ is a valid sample from the posterior probability density $p(\theta \mid D_{\rm obs})$ even though you never evaluated the likelihood. Now if you had an infinite amount of computing power and could draw a huge number of samples following this procedure and rejecting all samples that don't result in an exact match for your data then you could produce a posterior sampling for $p(\theta \mid D_{\rm obs})$ without ever evaluating the likelihood. Of course this method is absurd and intractable so ABC builds on this using two observations

1. This method will still be correct if sufficient statistics of the simulated data match the same sufficient statistics on the observed data.

2. The comparison between the simulation and the observation can be "softened" without significant loss of information.

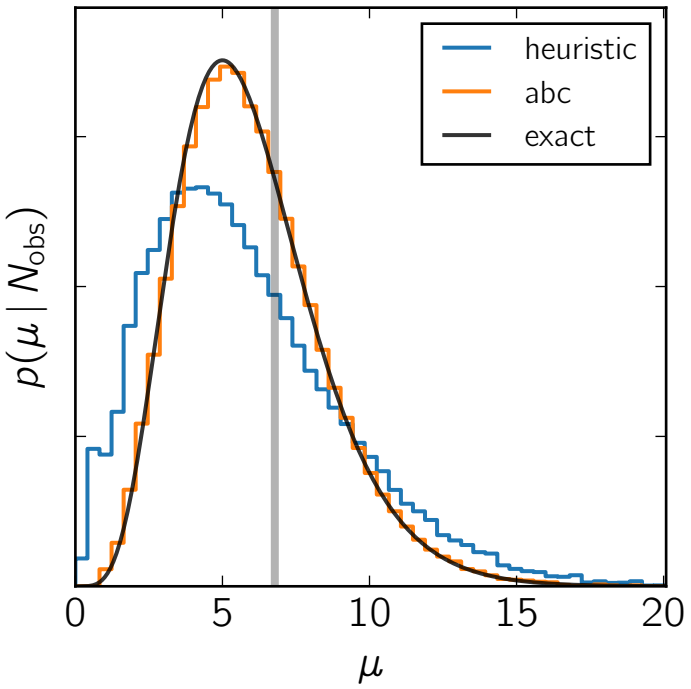

Figure 1. A comparison between a commonly used heuristic method for doing inference with stochastic models and ABC. The blue histogram shows the posterior inference for the rate parameter μ using the heuristic likelihood function from Equation (1), the orange histogram shows the ABC inference using the pseudo-likelihood from Equation (4), and the black curve shows the exact posterior from Equation (3). The vertical gray line shows the true value of μ that was used to generate the data.

The first point is correct by definition but we will return to the relevant subtleties below. The second point is less obvious.

- Qualitative justification.
- Theoretical basis.
- Methodological questions: (a) sufficient statistics, (b) distance function, (c) inference algorithm.

4. THE OCCURRENCE RATE OF SMALL AND COOL EXOPLANETS

- Sample selection (Burke et al. 2015).
- Generative model: power laws, discussion of Poisson process.
- Observation model: (Christiansen et al. 2013, 2015; Burke et al. 2015)
- Statistics, distance metric
- Inference method: PMC, MCMC, etc.
- Results: compare to Burke et al. (2015)

5. THE MULTIPLICITY & MUTUAL INCLINATION OF EXOPLANETARY SYSTEMS

- Sample selection: looks as a function of stellar type.
- Generative model: more flexible than power laws.
- Observation model: calibrating the completeness and caveats: CITE Christiansen note
- Inference method: PMC, MCMC, etc.
- Results

6. THE POTENTIAL IMPACT OF APPROXIMATE BAYESIAN COMPUTATION

- More general generative models means fewer and weaker assumptions.
- False positives.
- Physical models.

All of the code used in this project is available from `https://github.com/dfm/exoabc` under the MIT open-source software license. This code (plus some dependencies) can be run to re-generate all of the figures and results in this paper; this version of the paper was generated with git commit `8bbce60` (2016-11-09).

DFM would like to thank Jessi Cisewski for introducing him and the SAMSI exoplanet populations working group to the method of ABC. It is a pleasure to also thank Eric Agol, Eric Ford, David Hogg, and Richard Wilkinson, for helpful discussions and contributions.

DFM: Thank SAMSI.

This research made use of the NASA *Astrophysics Data System* and the NASA Exoplanet Archive. The Exoplanet Archive is operated by the California Institute of Technology, under contract with NASA under the Exoplanet Exploration Program.

This paper includes data collected by the *Kepler* mission. Funding for the *Kepler* mission is provided by the NASA Science Mission directorate. We are grateful to the entire *Kepler* team, past and present. Their tireless efforts were all essential to the tremendous success of the mission and the successes of *K2*, present and future.

These data were obtained from the Mikulski Archive for Space Telescopes (MAST). STScI is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS5-26555. Support for MAST is provided by the NASA Office of Space Science via grant NNX13AC07G and by other grants and contracts.

Computing resources were provided by High Performance Computing at New York University.

Facility: Kepler

Software: *corner.py* (Foreman-Mackey 2016), *matplotlib* (Hunter et al. 2007), *numpy* (Van Der Walt et al. 2011), *scipy* (Jones et al. 2001)

REFERENCES

Ballard, S., & Johnson, J. A. 2016, ApJ, 816, 66

Burke, C. J., Christiansen, J. L., Mullally, F., et al. 2015, ApJ, 809, 8

Christiansen, J. L., Clarke, B. D., Burke, C. J., et al. 2013, ApJS, 207, 35

—. 2015, ApJ, 810, 95

Fang, J., & Margot, J.-L. 2012, ApJ, 761, 92

Foreman-Mackey, D. 2016, The Journal of Open Source Software, 24, doi:10.21105/joss.00024

Hunter, J. D., et al. 2007, Computing in science and engineering, 9, 90

Jones, E., Oliphant, T., Peterson, P., et al. 2001, SciPy: Open source scientific tools for Python, ,

Tremaine, S., & Dong, S. 2012, AJ, 143, 94

Van Der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, Computing in Science & Engineering, 13, 22