Astronomy and Computing 13 (2015) 1–11

Contents lists available at ScienceDirect

Astronomy and Computing

journal homepage: www.elsevier.com/locate/ascom

Full length article

# COSMOABC: Likelihood-free inference via Population Monte Carlo Approximate Bayesian Computation

E.E.O. Ishida[a,*], S.D.P. Vitenti[b], M. Penna-Lima[c,d], J. Cisewski[e], R.S. de Souza[f], A.M.M. Trindade[g,h], E. Cameron[i], V.C. Busti[j,k]
for the COIN collaboration

[a] Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Straße 1, 85748, Garching, Germany
[b] $\mathcal{G}\mathbb{R}\varepsilon\mathbb{C}\mathcal{O}$ – Institut d'Astrophysique de Paris, UMR7095 CNRS, Université Pierre & Marie Curie, 98, bis boulevard Arago, 75014 Paris, France
[c] APC, AstroParticule et Cosmologie, Université Paris Diderot, UMR7164 CNRS/IN2P3, 10 rue Alice Domon et Léonie Duquet, 75013 Paris, France
[d] Instituto Nacional de Pesquisas Espaciais, Divisão de Astrofísica, Av. dos Astronautas 1758, 12227-010, São José dos Campos – SP, Brazil
[e] Department of Statistics, Yale University, New Haven, CT, 06511, USA
[f] MTA Eötvös University, EIRSA "Lendulet" Astrophysics Research Group, Budapest 1117, Hungary
[g] Instituto de Astrofísica e Ciências do Espaço, Universidade do Porto, CAUP, Rua das Estrelas, PT4150-762 Porto, Portugal
[h] Departamento de Física e Astronomia, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre 687, PT4169-007 Porto, Portugal
[i] Department of Zoology, University of Oxford, Tinbergen Building, South Parks Road, Oxford, OX1 3PS, United Kingdom
[j] Astronomy, Cosmology and Gravity Centre (ACGC), Department of Mathematics and Applied Mathematics, University of Cape Town, Rondebosch 7701, Cape Town, South Africa
[k] Departamento de Física Matemática, Instituto de Física, Universidade de São Paulo, CP 66318, CEP 05508-090, São Paulo - SP, Brazil

ARTICLE INFO

*Article history:*
Received 11 May 2015
Accepted 5 September 2015
Available online 16 September 2015

*Keywords:*
Galaxies: statistics
(cosmology:) large-scale structure of universe

ABSTRACT

Approximate Bayesian Computation (ABC) enables parameter inference for complex physical systems in cases where the true likelihood function is unknown, unavailable, or computationally too expensive. It relies on the forward simulation of mock data and comparison between observed and synthetic catalogues. Here we present COSMOABC, a Python ABC sampler featuring a *Population Monte Carlo* variation of the original ABC algorithm, which uses an adaptive importance sampling scheme. The code is very flexible and can be easily coupled to an external simulator, while allowing to incorporate arbitrary distance and prior functions. As an example of practical application, we coupled COSMOABC with the NUMCOSMO library and demonstrate how it can be used to estimate posterior probability distributions over cosmological parameters based on measurements of galaxy clusters number counts without computing the likelihood function. COSMOABC is published under the GPLv3 license on PyPI and GitHub and documentation is available at http://goo.gl/SmB8EX.



## 1. Introduction

The precision era of cosmology marks the transition from a data-deprived field to a data-driven science on which statistical methods play a central role. The ever-increasing data deluge must be tackled with new and innovative statistical methods in order to improve our understanding of the key ingredients driving our Universe (e.g., Borne, 2009; Ball and Brunner, 2010; de Souza et al., 2014; de Souza and Ciardi, 2015). Given the continuous inflow of new data, one does not start an analysis from scratch for every new telescope, but is guided by previous knowledge accumulated through experience. A new experiment provides extra information which needs to be incorporated into the larger picture, representing a small update on the previous body of knowledge. Such a learning process is a canonical scenario to be embedded in a Bayesian framework, which allow us to update our degree of belief on a set of model parameters[1] whenever new and independent data are acquired.

* Corresponding author.
*E-mail address:* emille@mpa-garching.mpg.de (E.E.O. Ishida).

[1] For the purposes of this work, we will only be interested in the *parameter values* of a given model. However, it is important to stress that in a completely Bayesian

http://dx.doi.org/10.1016/j.ascom.2015.09.001
2213-1337/

A standard Bayesian analysis specifies prior distributions on unknown parameters, defines which parameter values better describe the relationship between the model, the prior and the new data, and then finds the posterior distribution — either analytically or via sampling techniques, as e.g. with Markov Chain Monte Carlo (MCMC; Metropolis et al., 1953). This analysis requires a proper construction of the likelihood function, which is not always well known or easy to handle. A common solution would be to construct a model for the likelihood (e.g. a Gaussian) followed by MCMC, with the expectation that this hypothesis is not too far from the true. Nonetheless, the challenge of performing parameter inference from an unknown or intractable likelihood function is becoming familiar to the modern astronomer. Recent efforts to overcome observational selection biases in the study of massive (Sana et al., 2012) and not-so massive (Janson et al., 2014) stars, to account for windowing effects, errors and/or gaps in time-series of X-ray emission from active galactic nuclei (Uttley et al., 2002; Shimizu and Mushotzky, 2013) and UV emission from stellar coronae (Kashyap et al., 2002) have been reported.

The development of *ad hoc* approaches to this problem within astronomy have proceeded independently from their long history within the field of population genetics. The latter have ultimately been formalized into a rigorous statistical technique known as Approximate Bayesian Computation (ABC). The intuition of ABC dates back to a thought-experiment in Rubin (1984), where the basic ABC rejection sampler is used to illustrate Bayes Theorem. Tavaré et al. (1997) employs an acceptance–rejection method in the context of population genetics, while Pritchard et al. (1999) presents the first implementation of a basic ABC algorithm. Only recently has the ABC approach been introduced and applied to astronomical problems (Cameron and Pettitt, 2012; Schafer and Freeman, 2012; Weyant et al., 2013; Robin et al., 2014). This work is part of a larger endeavour natural consequence of such initial efforts. Following the philosophy behind the *Cosmostatistics Initiative* (COIN),[2] we present a tool which enables astronomers to easily introduce ABC techniques into their daily research.

The cornerstone of the ABC approach is our capability of performing quick and reliable computer simulations which mimic the observed data in the best possible way (this is called *forward simulation inference*). In this context, our task relies on performing a large number of simulations and quantifying the "distance" between the simulated and observed catalogues. The better a parametrization reproduces the observed data in a simulated context, the closer it is to the "true" model. From this simple reasoning, many alternatives were developed to optimize the parameter space sampling and the definition of distance. One of such examples is the work of Marjoram et al. (2003), who proposes a merger between the standard MCMC algorithm and the ABC rejection sampling. In astronomy, the method was used by Robin et al. (2014) to constrain the Milky Way thick disk formation. Going one step further, Beaumont et al. (2009a) propose to evolve an initial set of parameter values (or *particle system*) through incremental approximations to the true posterior distribution. The *Population Monte Carlo ABC* (PMC-ABC), method was used to make inferences on rate of morphological transformation of galaxies at high redshift (Cameron and Pettitt, 2012) and proved to be efficient in tracking the Hubble parameter evolution from type Ia supernova measurements, despite the contamination from type II supernova (Weyant et al., 2013). More recently, Lin and Kilbinger (2015) used ABC to predict weak lensing peak counts and Killedar et al. (2015) applied a weighted variant of the algorithm to cluster strong lensing cosmology.

This work introduces cosmoabc,[3] the first publicly available[4] Python ABC package for astronomy.[5] The package is structured so that the simulation, priors and distance functions are given as input to the main PMC-ABC sampler. In this context, users can easily connect the ABC algorithm to their own simulator and verify the effectiveness of the tool in their own astronomical problems. The package also contains exploratory tools which help defining a meaningful distance function and consequently point to appropriate choices before the sampler itself is initiated.

We first guide the user through a very simple toy model, in order to clarify how the algorithm and the package work. Subsequently, as an example of cosmological application, we show how the machinery can be used to define credible intervals over cosmological parameters based on galaxy clusters catalogues. Simulations for this example were performed using the *Numerical Cosmology* library (numcosmo; Vitenti and Penna-Lima, 2014). The connection between cosmoabc and numcosmo is implemented as an independent module that can be easily adapted to other cosmological probes.

The outline of this article is as follows. In Section 2, we give an overview of Bayesian perspective and the ABC algorithm. cosmoabc package is presented in Section 3 through a simple toy model. Section 4 describes in detail how to connect cosmoabc and numcosmo to obtain constrains over cosmological parameters from galaxy cluster number counts. Our final remarks are presented in Section 5.

## 2. Bayesian approaches to parameter inference

Statistical inference on unknown parameters is often a primary goal of a physical experiment design. Although it is possible, and at times even desirable, to encounter some unpredictable behaviour among the outcomes of an experiment or a measurement, in many situations the experimentation aims at establishing constraints over the parameters of a model. In other words, the desire is to use real-world data to check prevailing theories.

In the Bayesian framework, the data are seen as the accessible truth regarding a given physical process and the model as a representation of our understanding of such process. This approach is data-centred and allows us to update the model whenever new information becomes available. In other words, our goal is to determine the probability of a model given the data,

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}, \tag{1}$$

where $\boldsymbol{\theta}$ is the vector of model parameters, $\mathcal{D}$ the data set, $p(\boldsymbol{\theta}|\mathcal{D})$ is called the *posterior*, the prior, $p(\boldsymbol{\theta})$, represents our initial expectations towards the model and $p(\mathcal{D})$ is a normalization constant.

In this context, the model parameters themselves are considered random variables and each individual measurement corresponds to one realization of them. Thus, once our prior is confronted with the data, the outcome is a posterior probability distribution function (PDF). Using the posterior distributions we can determine *credible intervals*, which represent our uncertainty

approach all the elements and hypotheses forming the model can be considered part of the *prior*. In this sense, with the arrival of new essential information, the Bayesian approach allows for completely redefinition of the model itself (Kruschke, 2011).

2 http://goo.gl/rQZSAB.

3 https://pypi.python.org/pypi/CosmoABC.

4 Shortly after cosmoabc was released Akeret et al. (2015) also presented a Python package for forward modelling through PMC-ABC.

5 For similar tools in the context of biology and genetics, see e.g. Liepe et al. (2010); Oaks (2014).