HPC-accelerated Approximate Bayesian Computation for Biological Science

Ritabrata Dutta

Abstract Approximate Bayesian computation (ABC) provides us a rigorous tool to perform parameter inference for models without an easily accessible likelihood function. Here we give a short introduction to ABC, focusing on applications in biological science: estimation of parameters of an epidemiological spreading process on a network and a numerical platelets deposition model. Furthermore, we introduce users to a Python suite implementing ABC algorithms, with optimal use of high performance computing (HPC) facilities.

Key words: Approximate Bayesian Computation, Biological science, ABCpy

Introduction

With the recent innovations in biological science, we are increasingly facing large datasets of varied type and more realistic but complex models of natural phenomenon. This trend has led to a scenario where we do not easily have a likelihood function which is available in closed form and thus easy to evaluate at any given point (as required by most Monte Carlo and Markov chain Monte Carlo methods). Thus, traditional likelihood based inference, as Maximum likelihood or Bayesian methodology, is not possible. Still, if from the complex model, given values of the parameters that index it, we can forward simulate pseudo-dataset, a new methodology becomes available, namely Approximate Bayesian Computation (ABC). Models that have this possibility of forward simulation are known as simulator-based models and are becoming more and more popular in diverse fields of science [Martinez et al., 2016, Turchin et al., 2013, Schaye et al., 2015]; just restricting to the biological domain we can find many examples: evolution of genomes [Marttinen

Ritabrata Dutta
Institute of Computational Science, Università della Svizzera italiana, Switzerland, e-mail: duttar@usi.ch

et al., 2015], numerical model of platelet deposition [Chopard et al., 2017], demographic spread of a species [Excoffier et al., 2013] among many. Research in statistical science in the last decade or so, has illustrateted how ABC can be a tool to infer and calibrate the parameters of these models.

The fundamental rejection ABC sampling scheme iterates between three step: First a pseudo-dataset, $\boldsymbol{x}^{\text{sim}}$, is simulated from the simulator-based model $\mathscr{M}(\boldsymbol{\phi})$ for a fixed parameter value of $\boldsymbol{\phi}$. Then we compute a measure of the closeness between $\boldsymbol{x}^{\text{sim}}$ and $\boldsymbol{x}^0$, the observed dataset, using a pre-defined discrepancy measure $d(\boldsymbol{x}^{\text{sim}},\boldsymbol{x}^0)$. Finally, based on this discrepancy measure, ABC accepts the parameter value $\boldsymbol{\phi}$ when $d(\boldsymbol{x}^{\text{sim}},\boldsymbol{x}^0)$ is less than a pre-specified threshold value ε.

Following this ABC sampler, the intractable likelihood $\mathscr{L}(\boldsymbol{\phi})$ is approximated by $\mathscr{L}_{d,\varepsilon}(\boldsymbol{\phi})$ for some $\varepsilon > 0$, where

$$\mathscr{L}_{d,\varepsilon}(\boldsymbol{\phi}) \propto P(d(\boldsymbol{x}^{\text{sim}},\boldsymbol{x}^0) < \varepsilon) \tag{1}$$

and, as a consequence, the accepted parameters follow the posterior distribution of $\boldsymbol{\phi}$ conditional on $d(\boldsymbol{x}^{\text{sim}},\boldsymbol{x}^0) < \varepsilon$:

$$p_{d,\varepsilon}(\boldsymbol{\phi}|\boldsymbol{x}^0) \propto P(d(\boldsymbol{x}^{\text{sim}},\boldsymbol{x}^0) < \varepsilon)\pi(\boldsymbol{\phi}).$$

For a better approximation of the likelihood function, computationally efficient sequential ABC algorithms [Marin et al., 2012, Lenormand et al., 2013, Albert et al., 2015] decrease the value of the threshold ε adaptively while exploring the parameter space.

The crucial aspect for a good ABC approximation to the likelihood function is the choice of the summary statistics, as we define the discrepancy measure between $\boldsymbol{x}^{\text{sim}}$ and $\boldsymbol{x}^0$ through a distance between the extracted summary statistics from $\boldsymbol{x}^{\text{sim}}$ and $\boldsymbol{x}^0$. Knowledge domain driven summary statistics are normally chosen keeping in mind that we want to minimize the loss of information on $\boldsymbol{\phi}$ contained in the data through the choice of summary statistics. But one can also rely on automatic summary selection for ABC, thus removing a subjective component in this choice, as described in Fearnhead and Prangle [2012], Pudlo et al. [2015], Jiang et al. [2015] and Gutmann et al. [2017].

ABCpy

ABC provides a tool for statistical inference for simulator-based models, still, the necessity to simulate lots of pseudo-data, makes the algorithm extremely computationally expensive when data-simulation itself is costly. Further, the varied types of data sets available in different domain specific problems have hindered the applicability of ABC algorithms to many applied science domains. Recently, [Dutta et al., 2017a,d], have developed an High Performance Computing framework to efficiently

parallelize different ABC algorithms which we believe will be extremely beneficial for inferential problems across different scientific domains.

ABC and HPC were first brought together in the ABC-sysbio package for the systems biology community, where sequential Monte Carlo ABC (ABC-SMC) [T. Toni, 2009] algorithm was efficiently parallelized using graphics processing units (GPUs). The goal of ABCpy was to overcome the need for users to have knowledge of parallel programming, as is required for using ABC-sysbio [Liepe et al., 2010], and also to make a software package available for scientists across domains. These objectives were partly addressed by parallelization of ABC-SMC using MPI/OpenMPI Stram et al. [2015], and by making ABC-SMC available for the astronomical community Jennings and Madigan [2016]. Regardless of these advances, a recent ABC review article Lintusaari et al. [2017] highlights the depth and breadth of available ABC algorithms, which can be optimally efficient only via parallelization in an HPC environment Kulakova et al. [2016], Chiachio et al. [2014]. These developments emphasized the need of a generalized HPC supported platform for efficient ABC algorithms, which can be parallelized on multi-processor computers or computing clusters and is accessible to a broad range of scientists.

ABCpy addressed this need for an user-friendly scientific library of ABC algorithms, which is written in Python and designed in a highly modular fashion. Existing ABC software suites are mainly domain-specific and optimized for a narrower class of problems. Modularity of ABCpy makes it intuitive to use and easy to extend. Further, it enables users to run ABC sampling schemes in parallel without too much re-factoring of existing code. ABCpy includes likelihood free inference schemes, both based on discrepancy measures and approximate likelihood, providing a complete environment to develop new ABC algorithms.

Illustrative Applications

To highlight the versatility of ABC and ABCpy in diverse applied problems, we point the interested reader to two recent research papers with biological applications in mind: a) estimation of parameters of an epidemiological spreading process on a contact network[Dutta et al., 2017c] and b) estimation of parameters of a numerical platelets deposition model [Dutta et al., 2017b].

Epidemics on a Contact Network

Infectious diseases are studied to understand their spreading mechanisms, to evaluate control strategies and to predict the risk and course of future outbreaks. Because people only interact with a small number of individuals, and because the structure of these inter- actions matters for spreading processes, the pairwise relationships between individuals in a population can be usefully represented by a network. For

modeling the spread of infections on a human contact network, we consider a simple spreading process, i.e., the standard susceptible-infected (SI) process with unit infectivity on a fixed network [Zhou et al., 2006, Staples et al., 2016]. In this model, there are only two states, susceptible and infected, and this process is suitable for modeling the spread of pathogens in contact networks because a single successful exposure can be sufficient for transmission. In this process, at each time step, each infected node chooses one of its neighbors with equal probability regardless of their status (susceptible or infected), and if this neighboring node is susceptible, the node successfully infects it with probability θ. We denote this model by $\mathscr{M}_S$ and parametrize it in terms of the spreading rate θ and of the seed node (the node representing the first infected person) n_{SN}. For given values of these two parameters, $n_{\text{SN}} = n^*_{\text{SN}}$ and $\theta = \theta^*$, we can forward simulate the evolving epidemic over time using the $\mathscr{M}_S$ model as

$$\mathscr{M}_S[n^*_{\text{SN}}, \theta^*] \rightarrow \{\mathbb{N}_{\mathbb{I}}(t), t = 0, \ldots, T\}, \tag{2}$$

where $\mathbb{N}_{\mathbb{I}}(t)$ is a list of infected nodes at time t. We simulated an epidemic of a disease using the above simple contagion process in an Indian village contact network[Banerjee et al., 2013]. The network has 354 nodes and 1541 edges, representing 354 villagers and reported contacts and social relationships among them. The epidemic is simulated using $\theta^0 = 0.3$, $n^0_{\text{SN}} = 70$, and the observed dataset $\boldsymbol{x}^0$ is the infected nodes $\mathbb{N}_{\mathbb{I}}(t)$ for $t = t_0, \ldots, T$ with $t_0 = 20$ and $T = 70$. The marginal posterior distributions and the Bayes estimates of (θ, n_{SN}) are show in Figure 1. The inferred posterior distributions for the epidemics on the Indian village contact network, is concentrated around the true parameter values. The Bayes estimates are also in a very small neighborhood of the true value, specifically the estimated seed-node $(\hat{n}_{\text{SN}})$ has a shortest path distance of 1 from n^0_{SN} in both the cases.

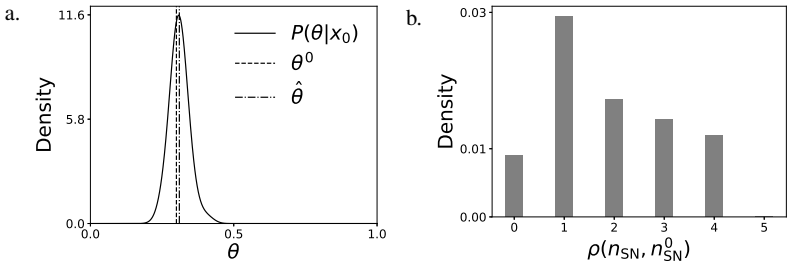

Fig. 1: **Simple contagion model on Indian village contact network.** Panel **a** shows the density of the inferred marginal posterior distribution and Bayes estimate of θ, given $\boldsymbol{x}^0$, the epidemics on the Indian village contact network. Panel **b** displays the average marginal posterior distribution at different distances from the true seed-node n^0_{SN}. The shortest path length distance between $n^0_{\text{SN}} = 70$ and $\hat{n}_{\text{SN}} = 59$ is 1.

For details on how the inference was performed via ABC and ABCpy, we direct readers to Dutta et al. [2017]. We can further extend this inferential approach to any

complex spreading processes on a network, e.g. inference of parameters of complex contagion model representing a disinformation campaign on a social network is reported in Dutta et al. [2017].

Platelet Deposition Model

Chopard et al. [2015], Chopard et al. [2017] has recently developed a numerical model that quantitatively describes how platelets in a shear flow adhere and aggregate on a deposition surface. Five parameters specify the deposition process and are relevant for a biomedical understanding of the phenomena. An experimental observations can be collected from a patient, at time intervals, on the average size of the aggregation clusters, their number per mm^2, the number of platelets and the ones activated per $\mu\ell$ still in suspension. In Dutta et al. [2017b], we have demonstrated that approximate Bayesian computation (ABC) can be used to automatically explore the parameter space of this numerical model. To illustrate the performance of ABC, in Figure 2, we show the inferred posterior distribution of the parameters (adhesion rate p_{Ad}, the aggregation rates p_{Ag} and p_T, the deposition rate of albumin p_F, and the attenuation factor a_T) of the platelet deposition model. For details on the specific model and on how the inference was performed via ABC, we direct readers to Chopard et al. [2017] and Dutta et al. [2017b] correspondingly.

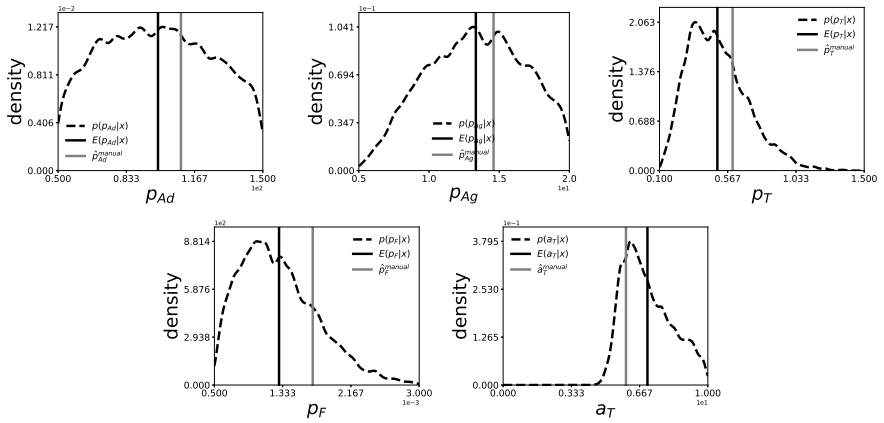

Fig. 2: Marginal posterior distribution (black-dashed) and Bayes Estimate (back-solid) of $(p_{Ad}, p_{Ag}, p_T, p_F, a_T)$ for collective data set generated from of 7 patients. The smoothed marginal distribution is created by a Gaussian-kernel density estimator on 5000 samples drawn from the posterior distribution using Simulated annealing approximate Bayesian computation [Albert et al., 2015]. The (gray-solid) line indicates the manually estimated values of the parameters as in [Chopard et al., 2017].

The proposed approach can be applied patient per patient, in a systematic way, without the bias of a human operator. In addition, the approach is computationally fast enough to provide results in an acceptable time for contributing to a new medical diagnosis, by giving data that no other known method can provide.

Conclusion

We would like to stress here the fact that ABC inference scheme provides not only a point estimate of the parameters of interest but also their entire (approximated) posterior distribution thus allowing for uncertainty quantification: the higher the variability of the posterior distribution the higher the uncertainty inherent in the inferential scheme. Via the ABC approximated posterior one can then construct credible intervals and perform hypothesis testing. Furthermore ABC allows to compare possible alternative models by simply adding, to the three steps Rejection ABC scheme illustrated above, an additional initial layer where first a model index is sampled from the model prior distribution and then, once a model has been selected a regular ABC scheme within that model is performed. For details on ABC model selection via random forest approach see Pudlo et al. [2015].

Acknowledgements The research was supported by Swiss National Science Foundation Grant No. 105218 163196 (Statistical Inference on Large-Scale Mechanistic Network Models). We also thank Dr. Marcel Schoengens, CSCS, ETH Zurich for helps regarding HPC and Swiss National Super Computing Center for providing computing resources.

References

Carlo Albert, Hans R. Künsch, and Andreas Scheidegger. A simulated annealing approach to approximate Bayesian computations. *Statistics and Computing*, 25: 1217–1232, 2015.

Abhijit Banerjee, Arun G Chandrasekhar, Esther Duflo, and Matthew O Jackson. The diffusion of microfinance. *Science*, 341(6144):1236498, 2013.

Manuel Chiachio, James L. Beck, Juan Chiachio, and Guillermo Rus. Approximate Bayesian computation by subset simulation. *SIAM J. Sci. Comput.*, 36(3):A1339–A1358, 2014.

B. Chopard, D. Ribeiro de Sousa, J. Latt, F. Dubois, C. Yourassowsky, P. Van Antwerpen, O. Eker, L. Vanhamme, D. Perez-Morga, G. Courbebaisse, and K. Zouaoui Boudjeltia. A physical description of the adhesion and aggregation of platelets. *ArXiv e-prints*, 2015.

Bastien Chopard, Daniel Ribeiro de Sousa, Jonas Lätt, Lampros Mountrakis, Frank Dubois, Catherine Yourassowsky, Pierre Van Antwerpen, Omer Eker, Luc Van-

hamme, David Perez-Morga, et al. A physical description of the adhesion and aggregation of platelets. *Royal Society Open Science*, 4(4):170219, 2017.

R. Dutta, A. Mira, and J.-P. Onnela. Bayesian Inference of Spreading Processes on Networks. *ArXiv e-prints*, September 2017.

R Dutta, M Schoengens, J.P. Onnela, and Antonietta Mira. ABCpy: A user-friendly, extensible, and parallel library for approximate Bayesian computation. In *Proceedings of the Platform for Advanced Scientific Computing Conference*. ACM, June 2017a.

Ritabrata Dutta, Bastien Chopard, Jonas Lätt, Frank Dubois, Karim Zouaoui Boudjeltia, and Antonietta Mira. Parameter estimation of platelets deposition: Approximate bayesian computation with high performance computing. *arXiv preprint arXiv:1710.01054*, 2017b.

Ritabrata Dutta, Antonietta Mira, and Jukka-Pekka Onnela. Bayesian inference of spreading processes on network. *arXiv preprint arXiv:1709.08862*, 2017c.

Ritabrata Dutta, Marcel Schoengens, Avinash Ummadisingu, Jukka-Pekka Onnela, and Antonietta Mira. Abcpy: A high-performance computing perspective to approximate bayesian computation. *arXiv preprint arXiv:1711.04694*, 2017d.

Laurent Excoffier, Isabelle Dupanloup, Emilia Huerta-Sánchez, Vitor C Sousa, and Matthieu Foll. Robust demographic inference from genomic and snp data. *PLoS genetics*, 9(10):e1003905, 2013.

Paul Fearnhead and Dennis Prangle. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74 (3):419–474, 2012.

Michael U Gutmann, Ritabrata Dutta, Samuel Kaski, and Jukka Corander. Likelihood-free inference via classification. *Statistics and Computing*, pages 1–15, 2017.

E. Jennings and M. Madigan. astroABC: An Approximate Bayesian Computation Sequential Monte Carlo sampler for cosmological parameter estimation. *ArXiv:1608.07606*, 2016.

Bai Jiang, Tung-yu Wu, Charles Zheng, and Wing H Wong. Learning summary statistic for approximate Bayesian computation via deep neural network. *arXiv preprint arXiv:1510.02175*, 2015.

Lina Kulakova, Panagiotis Angelikopoulos, Panagiotis E. Hadjidoukas, Costas Papadimitriou, and Petros Koumoutsakos. Approximate Bayesian computation for granular and molecular dynamics simulations. In *Proceedings of the Platform for Advanced Scientific Computing Conference*, PASC '16, pages 4:1–4:12. ACM, 2016. doi: 10.1145/2929908.2929918.

Maxime Lenormand, Franck Jabot, and Guillaume Deffuant. Adaptive approximate bayesian computation for complex models. *Computational Statistics*, 28 (6):2777–2796, 2013.

Juliane Liepe, Chris Barnes, Erika Cule, Kamil Erguler, Paul Kirk, Tina Toni, and Michael P.H. Stumpf. ABC-SysBio – approximate Bayesian computation in Python with GPU support. *Bioinformatics*, 26(14):1797–1799, 2010. doi: 10.1093/bioinformatics/btq278.

Jarno Lintusaari, Michael U Gutmann, Ritabrata Dutta, Samuel Kaski, and Jukka Corander. Fundamentals and recent developments in approximate Bayesian computation. *Systematic Biology*, 66(1):e66–e82, 2017. doi: 10.1093/sysbio/syw077. URL https://doi.org/10.1093/sysbio/syw077.

Jean-Michel Marin, Pierre Pudlo, ChristianP. Robert, and RobinJ. Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22(6): 1167–1180, 2012. ISSN 0960-3174. doi: 10.1007/s11222-011-9288-2. URL http://dx.doi.org/10.1007/s11222-011-9288-2.

Esteban A. Martinez, Christine A. Muschik, Philipp Schindler, Daniel Nigg, Alexander Erhard, Markus Heyl, Philipp Hauke, Marcello Dalmonte, Thomas Monz, Peter Zoller, and Rainer Blatt. Real-time dynamics of lattice gauge theories with a few-qubit quantum computer. *Nature*, 534(7608):516–519, 2016. doi: 10.1038/nature18318.

Pekka Marttinen, Nicholas J Croucher, Michael U Gutmann, Jukka Corander, and William P Hanage. Recombination produces coherent bacterial species clusters in both core and accessory genomes. *Microbial Genomics*, 1(5), 2015.

Pierre Pudlo, Jean-Michel Marin, Arnaud Estoup, Jean-Marie Cornuet, Mathieu Gautier, and Christian P Robert. Reliable ABC model choice via random forests. *Bioinformatics*, 32(6):859–866, 2015.

Joop Schaye, Robert A. Crain, Richard G. Bower, Michelle Furlong, Matthieu Schaller, Tom Theuns, Claudio Dalla Vecchia, Carlos S. Frenk, I. G. McCarthy, John C. Helly, Adrian Jenkins, Y. M. Rosas-Guevara, Simon D. M. White, Maarten Baes, C. M. Booth, Peter Camps, Julio F. Navarro, Yan Qu, Alireza Rahmati, Till Sawala, Peter A. Thomas, and James Trayford. The EAGLE project: simulating the evolution and assembly of galaxies and their environments. *Monthly Notices of the Royal Astronomical Society*, 446(1):521–554, 2015. doi: 10.1093/mnras/stu2058.

Patrick Staples, Mélanie Prague, De Gruttola Victor, and Jukka-Pekka Onnela. Leveraging contact network information in clustered randomized trials of infectious processes. *arXiv preprint arXiv:1610.00039*, 2016. URL https://arxiv.org/abs/1610.00039.

Alexander H. Stram, Paul Marjoram, and Gary K. Chen. al3c: high-performance software for parameter inference using Approximate Bayesian Computation. *Bioinformatics*, 31(21):3549–3551, 2015. doi: 10.1093/bioinformatics/btv393.

M. S. T. Toni. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 31(6):187–202, 2009.

Peter Turchin, Thomas E. Currie, Edward A. L. Turner, and Sergey Gavrilets. War, space, and the evolution of old world complex societies. *Proceedings of the National Academy of Sciences*, 110(41):16384–16389, 2013. doi: 10.1073/pnas.1308825110.

Tao Zhou, Jian-Guo Liu, Wen-Jie Bai, Guanrong Chen, and Bing-Hong Wang. Behaviors of susceptible-infected epidemics on scale-free networks with identical infectivity. *Physical Review E*, 74(5):056109, 2006. doi: 10.1103/PhysRevE.74.056109. URL https://doi.org/10.1103/PhysRevE.74.056109.