ABC and distributional random forests

Khanh Dinh, Simon Tavaré & Zijin Xiang

February 20, 2024

1 Introduction

Statistical inference for stochastic processes is often challenging because of the difficulty (or impossibility) of computing the likelihood function. In response to this challenge, population geneticists introduced the idea of ABC – Approximate Bayesian Computation (Tavaré *et al.*, 1997; Pritchard *et al.*, 1999). This approach exploited simulation to generate observations from the process of interest for parameter values θ generated from the prior. The resultant data are summarised and the summary statistics, S_{sim}, compared to the value of the observed summary, S_{obs}. Values of θ for which S_{sim} are close to S_{obs} are taken as observations from the (approximate) posterior of θ. Since its introduction ABC has flourished as part of the statisticians toolkit, a plethora of methods being discussed in Sisson *et al.* (2019) for example.

Among the challenges of ABC are choice of summary statistics, choice of metric to compare them, and implementation of some of the methods. This situation changed with the appearance of ABC-RF, a random forest method, in Raynal *et al.* (2019). The idea is to exploit random forest regression or classification to infer posterior features of the parameters or to compare hypotheses, respectively. The approach is implemented in the R packages `abcrf` and `DIYABC` and can provide estimates of the marginal posterior distribution of functions of each parameter, while avoiding issues such as choice of summary statistics and metric.

2 Distributional random forests

The principal drawback of ABC-RF is the lack of a simple way to study the *joint* posterior behavior of the parameters, which is necessary for studying goodness-of-fit through posterior predictive analysis, for example. The purpose of this note is to highlight the appearance of *distributional random forests* in (Cevid *et al.*, 2022), a method to address regression problems with multi-dimensional dependent variables. This approach, ABC-DRF, can readily be adapted for use in ABC to deal with full posterior distributions.

3 A toy model

We illustrate the approach with one example from Raynal *et al.* (2019), for which the marginal posteriors are known explicitly. The parameter is $\theta = (\theta_1, \theta_2)$, and the hierarchical model has $Y_i \mid \theta \sim \mathrm{N}(\theta_1, \theta_2)$, the normal distribution with mean θ_1 and variance θ_2; $\theta_1 \mid \theta_2 \sim \mathrm{N}(0, \theta_2)$, and finally $\theta_2 \sim \mathrm{IG}(4, 3)$, where $\mathrm{IG}(\alpha, \beta)$ denotes the inverse gamma distribution with density

$$f(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}(1/x)^{\alpha+1} \exp\left(-\beta/x\right), \quad x > 0.$$

We take a conditionally independent sample $Y = (Y_1, \ldots, Y_n)$ of size n with given values of θ, and compute

$$\bar{Y} = n^{-1} \sum Y_i, \quad S^2 = \sum (Y_i - \bar{Y})^2, \quad B = \frac{1}{2}(S^2 + 6 + n\bar{Y}^2/(n+1))$$

The posteriors of θ_1 and θ_2 given Y are dependent, but uncorrelated. Their joint law has

$$\theta_2 \mid Y \sim \mathrm{IG}\left(\frac{n}{2} + 4,\ B\right) \tag{1}$$

and

$$\theta_1 \mid \theta_2, Y \sim \mathrm{N}\left(\frac{n\bar{Y}}{n+1}, \frac{2\theta_2}{n+1}\right) \tag{2}$$

The marginal posterior of θ_1 is given in (Raynal *et al.*, 2019) as

$$\theta_1 \mid Y \sim \mu + \tau\, T_{n+8},$$

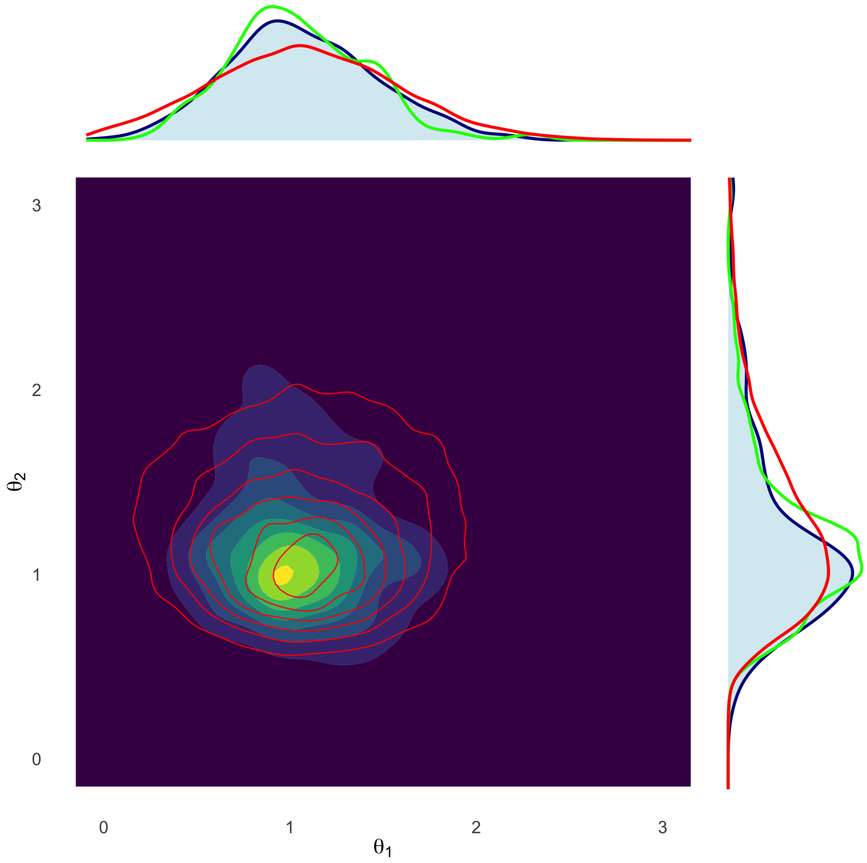
where

$$\mu = \frac{n\bar{Y}}{(n+1)}, \quad \tau^2 = \frac{2B}{(n+1)(n+8)}.$$

and T_m denotes the t-distribution with m degrees of freedom.

3.1 Example

We illustrate the ABC-RF and ABC-DRF methods for the toy model. We generated a reference set of 10,000 examples simulated from the model with $\alpha = 4, \beta = 3$, and chose the same set of summary statistics as in Raynal *et al.* (2019), resulting in a total of 61 features, 50 of which were U(0,1) noise and 11 related to the model. In the figure below we have plotted the marginal posteriors of θ_1 and θ_2 from ABC-RF (green lines), and the corresponding marginals inferred from ABC-DRF using the R implementation in `drf` (blue lines). The contour plot for the inferred joint posterior is in the body of the plot, and the true density contours, readily computable using (1) and (2), are superimposed (red lines). The fits seem convincing.



The results suggest that ABC-DRF will be a very useful addition to the statistician's toolbox.

References

Cevid, D., L. Michel, J. Näf, P. Bühlmann, and N. Meinshausen, 2022 Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. Journal of Machine Learning Research **23**: 1–79.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman, 1999 Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol Biol Evol **16**: 1791–1798.

Raynal, L., M. J-L., P. Pudlo, M. Ribatet, C. P. Robert, and A. Estoup, 2019 ABC random forests for Bayesian parameter inference. Bioinformatics **35**: 1720 – 1728.

Sisson, S. A., Y. Fan, and M. A. Beaumont, editors, 2019 *Handbook of Approximate Bayesian Computation*. Chapman & Hall/CRC.

Tavaré, S., D. J. Balding, R. C. Griffiths, and P. Donnelly, 1997 Inferring coalescence times from DNA sequence data. Genetics **145**: 505–518.