Stochastic Simulations

Autumn Semester 2023

Prof. Fabio Nobile *Assistant: Matteo Raviola*

Project 1

submission deadline: 16 January 2024

Approximate Bayesian Computation MCMC

1 Introduction and background

Many problems in data science involve estimating a set of parameters $\theta \in \Theta$ of a model $\mathcal{M}$ that describes the processes underlying the problem of interest. In the Bayesian inference paradigm, the uncertainty over such a set of parameters is quantified by means of a *posterior* distribution, which is often described through sampling techniques. In this context, one assumes the parameters to follow a *prior* distribution $\pi(\theta)$, which describes the current believe on the problem at hand, e.g., based on available expert knowledge. After having observed some data $\mathcal{D}$, this prior believe is updated by means of the *likelihood* function $\mathbb{P}(\mathcal{D}|\theta)$, which describes the plausibility of having generated such data under all possible different values of θ. The posterior distribution of interest, $f(\theta|\mathcal{D})$, is then determined by Bayes' rule

$$f(\theta|\mathcal{D}) = \mathbb{P}(\mathcal{D}|\theta)\pi(\theta)/\mathbb{P}(\mathcal{D}), \tag{1}$$

where $\mathbb{P}(\mathcal{D}) = \int_\Theta \mathbb{P}(\mathcal{D}|\theta)\pi(\theta)d\theta$, called the evidence, represents the normalizing constant.

Stochastic simulation approaches for generating observations from the posterior distribution $f(\theta|\mathcal{D})$ often depend on knowing explicitly the likelihood function $\mathbb{P}(\mathcal{D}|\theta)$, possibly up to a multiplicative constant (i.e. being able to evaluate it for any θ and $\mathcal{D}$). However, for many complex probabilistic models, such likelihoods are either inaccessible or computationally prohibitive to evaluate, so one has to resort to the so-called *likelihood-free* methods, of which, most notably the *Approximate Bayesian Computation* (ABC).

The simplest ABC rejection algorithm scheme can be found in Algorithm 1 below. In a nutshell, ABC algorithms sample candidate parameters $\theta^\star$ from the prior distribution $\pi(\theta)$, generate a data sample $\mathcal{D}^\star$ given the candidate parameters $\theta^\star$ and then compare it with the observed data $\mathcal{D}$ according to some pre-defined discrepancy metric $\rho(\cdot,\cdot)$ and tolerance ε. In the acceptance/rejection step, the candidate parameters are accepted as samples from the posterior if the simulated data is similar enough (in terms of the chosen $\rho(\cdot,\cdot)$ and ε) to the observed data $\mathcal{D}$. Otherwise, a new candidate is sampled from the prior. This scheme is then repeated until a (sufficiently large) sample of size N, distributed approximately as the posterior, is obtained. This approach requires suitable choices of the metric ρ and tolerance ε. Note that in fact, Algorithm 1 provides samples from the approximated posterior $f(\theta|\rho(\mathcal{D}^\star,\mathcal{D}) < \varepsilon)$. As $\varepsilon \to \infty$ it generates observations from the prior. If, on the contrary, $\varepsilon = 0$, and the observation $\mathcal{D}^\star$ is accepted only if $\mathcal{D}^\star = \mathcal{D}$, then the accepted observations come from the true posterior density $f(\theta|\mathcal{D})$. The choice of ε therefore reflects a tension between computability and accuracy of the method.

Algorithm 1 Basic ABC rejection algorithm

for $i = 1, \ldots, N$ **do**
  Sample candidate parameters from the prior distribution $\theta^* \sim \pi(\cdot)$
  Generate data from the underlying model given $\theta^\star$, $\mathcal{D}^* \sim \mathbb{P}(\cdot|\theta^*)$
  if $\rho(\mathcal{D}^\star, \mathcal{D}) < \varepsilon$ **then**
    Set $\theta^i \leftarrow \theta^\star$
  else
    Go back to Step 2.
  end if
end for

Remark 1.1. Oftentimes the comparison step is performed using a certain summary statistics $S(\cdot)$ from the observed and generated data, that is, one considers $\rho(S(\mathcal{D}^\star), S(\mathcal{D}))$. Such summary statistics can be the sample mean, the sample median, or a specific quantile, among others. This procedure is meant to alleviate high rejection rates and to lower the dimensionality in scenarios where $\mathcal{D}$ is high- or infinite-dimensional.

1.1 MCMC for reducing ABC rejection

The sampling strategy outlined in Algorithm 1 poses challenges on its own, even for simple models as we will see below. Acceptance rates can be very low as candidate parameter vectors are generated from the prior $\pi(\theta)$, which may be quite different from the posterior. Thus, several accelerating techniques have been proposed, for instance, to embed the ABC scheme within the well known Metropolis-Hastings framework as shown in Algorithm 2.

Algorithm 2 ABC-MCMC algorithm

Initialize θ_0
for $i = 0, \ldots, N$ **do**
  Sample candidate parameters θ^* from a proposal transition density q, $\theta^* \sim q(\theta_i, \cdot)$
  Generate data from the underlying model given $\theta^\star$, $\mathcal{D}^* \sim \mathbb{P}(\mathcal{D}|\theta^*)$
  if $\rho(\mathcal{D}^\star, \mathcal{D}) < \varepsilon$ **then**
    Set $\theta_{i+1} \leftarrow \theta^\star$ with probability $\alpha = \min\left(1, \frac{\pi(\theta^\star)q(\theta^\star, \theta_i)}{\pi(\theta_i)q(\theta_i, \theta^\star)}\right)$ and $\theta_{i+1} \leftarrow \theta_i$ otherwise
  else
    Set $\theta_{i+1} \leftarrow \theta_i$
  end if
end for

In this setting, Algorithm 2 generates a sequence of serially correlated samples from $f(\theta|\rho(\mathcal{D}^\star, \mathcal{D}) < \varepsilon)$. Determination of the chain length, N, is therefore obtained through a careful assessment of convergence of the chain and considerations of the chain's ability to explore the parameter space Θ (i.e., chain mixing).

Notice that the candidate vector of parameters is generated from an arbitrary proposal transition density $q(\cdot, \cdot)$ and accepted with a Metropolis–Hastings type acceptance probability, in which, however, the (intractable) likelihood ratio $\mathbb{P}(\cdot|\theta^*)/\mathbb{P}(\cdot|\theta_i)$ is coarsely approximated by 1 under the assumption that the simulated and observed data are sufficiently close according to the chosen metric $\rho(\cdot, \cdot)$, and 0 otherwise.

1.2 A model in pharmacokinetics

We want to apply the ideas described above to a dynamical model of the pharmacokinetics of Theophylline, a drug used in the treatment of asthma and chronic obstructive pulmonary disease. In pharmacokinetics one aims to study a drug of interest by describing its absorption, distribution, metabolism, and excretion mechanisms from the body. A fundamental concept in pharmacokinetics is drug clearance, that is, elimination of drugs from the body. Let X_t be the level of Theophylline concentration in blood at time t, then the evolution of X_t over time can be modeled by means of the following stochastic differential equation

$$dX_t = \left(\frac{DK_aK_e}{Cl} e^{-K_a t} - K_e X_t \right) dt + \sigma dW_t, \tag{2}$$

where D is the known drug oral dose received by a subject, K_e is the elimination rate constant, K_a the absorption rate constant, Cl the clearance of the drug, and σ the intensity of intrinsic stochastic noise driven by the Brownian motion W_t. The experimental design for a single hypothetical subject considers nine blood samples taken at $\{t_1, \ldots, t_9\} = \{0.25, 0.5, 1, 2, 3.5, 5, 7, 9, 12\}$ hours after dosing. The drug oral dose is chosen to be $D = 4$ mg and is administered starting from $t_0 = 0^+$. The initial drug concentration in blood is $X_0 = 0$. Inference is based on data $\{x_i\}_{i=1}^9$ collected at times $\{t_1, \ldots, t_9\}$ and the parameters of interest are $\theta = (K_e, K_a, Cl, \sigma)$, for which the following priors are considered: $\log K_e \sim \mathcal{N}(-2.7, 0.6^2)$, $\log K_a \sim \mathcal{N}(0.14, 0.4^2)$, $\log Cl \sim \mathcal{N}(-3, 0.8^2)$, $\log \sigma \sim \mathcal{N}(-1.1, 0.3^2)$.

2 Goals of the project

1. We first consider an academic example where the likelihood function and true posterior distribution f are actually known, so that we can compare the result of the ABC sampling with the true posterior.

   In this example, the observed data $\mathcal{D} = \{x_i\}_{i=1}^M \subset \mathbb{R}$ is an iid sample drawn with probability 1/2 from $\mathcal{N}(\theta, \sigma_1^2)$ and with probability 1/2 from $\mathcal{N}(\theta + a, \sigma_1^2)$. As a prior we take $\pi = \mathcal{N}(0, \sigma^2)$. Then, the posterior distribution is a Gaussian mixture given by

   $$f(\theta|\mathcal{D}) = \alpha \mathcal{N}\left(\frac{\sigma^2}{\sigma^2 + \sigma_1^2/M} \bar{x}, \frac{\sigma_1^2}{M + \sigma_1^2/\sigma^2} \right) + (1-\alpha) \mathcal{N}\left(\frac{\sigma^2}{\sigma^2 + \sigma_1^2/M} (\bar{x} - a), \frac{\sigma_1^2}{M + \sigma_1^2/\sigma^2} \right), \tag{3}$$

   with

   $$\alpha = \frac{1}{1 + \exp\left\{ a\left(\bar{x} - \frac{a}{2}\right) \frac{M}{M\sigma^2 + \sigma_1^2} \right\}}$$

   where $\bar{x} = \frac{1}{M} \sum_{i=1}^M x_i$ denotes the sample mean of the data and $\mathcal{N}\left(\mu, \sigma^2\right)$ denotes the density of a Gaussian random variable with mean μ and variance σ^2. We consider the following parameters: $M = 100$, $\sigma_1^2 = 0.1$, $\sigma^2 = 3$, $a = 1$ and we assume that the sample mean of the data is exactly $\bar{x} = 0$.

   Implement the basic ABC rejection algorithm described in Algorithm 1 for different tolerances $\varepsilon = \{0.75, 0.25, 0.1, 0.025\}$, until $N = 500$ samples are accepted, with discrepancy metric defined as

   $$\rho(S(\mathcal{D}^\star), S(\mathcal{D})) = |\bar{x}^* - \bar{x}| \tag{4}$$

where $\bar{x}^* = \frac{1}{M}\sum_{i=1}^{M} x_i^*$ is the sample mean of the *generated* data $\mathcal{D}^* = \{x_i^*\}_{i=1}^M$ according to the mechanism described above.

Report the acceptance rates observed for each of the considered tolerances. Plot the histogram of your samples along with the true mixture distribution in (3) for the 4 considered tolerances and comment on the results.

2. Prove that Algorithm 1 generates samples distributed as $f(\theta|\rho(\mathcal{D}^\star, \mathcal{D}) < \varepsilon)$ and, similarly, that $f(\theta|\rho(\mathcal{D}^\star, \mathcal{D}) < \varepsilon)$ is the stationary distribution of the chain generated by Algorithm 2.

3. Implement the ABC-MCMC method described in Algorithm 2, with random walk proposal $q(\theta, \cdot) = \mathcal{N}(\theta, \nu^2)$ and initial state $\theta_0 = 0$. Try different values of ν^2. Run the chain long enough to have an effective sample size $N_{\text{eff}} \approx 500$. Compare then your results with those obtained by the crude rejection ABC algorithm, in terms of acceptance rate and approximation of the posterior distribution.

4. Consider now the pharmacokinetic model (2). Generate synthetically the data $\mathcal{D} = \{x_i\}_{i=1}^9$ by simulating the model (2) with parameters $\theta = (0.08, 1.5, 0.04, 0.2)$ and recording the solution at the sampling times $\{t_1, \dots, t_9\}$ to get $n = 9$ values for the process X_t. (The process (2) is Gaussian and can be simulated exactly, but as an alterantive you can also use an Euler-Maruyama discretization with sufficiently small time step). Then, to construct a summary statistics $S(\mathcal{D})$, fit the following multivariate linear regression model

$$\theta = \beta_0 + \beta_1 x_1 + \cdots + \beta_9 x_9 + \xi \tag{5}$$

where $\theta = (\theta_1, \dots, \theta_4)$ is the vector of parameters, $\beta_i \in \mathbb{R}^4$, $i = 0, \dots, 9$, are unknown regression coefficients, $\{x_i\}_{i=1}^9$ are the generated data $\mathcal{D}$ and $\xi = (\xi_1, \dots, \xi_4)$ is a random vector with zero mean, independent components and constant variance. You can generate multiple values of the parameters $\theta^{(1)}, \dots, \theta^{(p)}$ from the prior distribution, and generate corresponding data $\mathcal{D}^{(1)}, \dots, \mathcal{D}^{(p)}$ to build the linear regression model. If $\hat{\beta}_i$, $i = 0, \dots, 9$, denote the estimated regression coefficients, then the summary statistics reads

$$S(\mathcal{D}) = \mathbb{E}(\theta|\mathcal{D}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_9 x_9 \in \mathbb{R}^4.$$

5. For tolerances $\varepsilon = \{0.25, 0.7, 1\}$, run $N = 10,000$ iterations of the ABC-MCMC algorithm (Algorithm 2) with initial state $\theta_0 = (0.07, 1.15, 0.05, 0.33)$ and discrepancy metric $\rho(S(\mathcal{D}^\star), S(\mathcal{D})) = \|S(\mathcal{D}^\star) - S(\mathcal{D})\|$, where $\|\theta\|^2 = \sum \frac{\theta_i^2}{(\theta_0)_i^2}$ is a weighted euclidean norm in $\mathbb{R}^4$. Choose, in particular, a suitable proposal for the MCMC and motivate your choice. Discuss the influence of the tolerance on the acceptance rate and the obtained posterior distribution. (You could look, for instance, at the marginal distribution of each of the four parameters). Provide an estimate for the posterior mean θ^{PM} of the four parameters.

6. Using the posterior mean computed at the previous point in the model (2), propose an efficient Monte Carlo estimator (eventually using variance reduction techniques) to estimate $\mathbb{E}[X_9]$, the expected concentration of Theophylline after 12 hours.

References

[1] Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavaré. Markov chain monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 100(26):15324–15328, 2003.

[2] Umberto Picchini. Inference for SDE models via approximate Bayesian computation. *Journal of Computational and Graphical Statistics*, 23(4):1080–1100, 2014.

[3] S. A. Sisson, Y. Fan, and Mark M. Tanaka. Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences*, 104(6):1760–1765, 2007.

[4] Scott A Sisson, Yanan Fan, and Mark Beaumont. *Handbook of approximate Bayesian computation*. CRC Press, 2018.