`abc-sde`: a Matlab package for ABC in stochastic differential equation models

https://sourceforge.net/projects/abc-sde/

Umberto Picchini
Centre for Mathematical Sciences, Lund University, Sweden

LUND UNIVERSITY

A note of caution

Notice at the moment the `abc-sde` package is dependent on the Matlab Statistics Toolbox (dependence "soon" to be removed).

Motivation

We want to enable inference for parameters of SDE models via approximate Bayesian computation (ABC). The observational framework is quite general:

- one- or multi-dimensional SDEs;
- system states $X_t \in \mathbb{R}^d$ are observed with error;
- some states might be unobserved (partially observed system);

$$\begin{cases} dX_t = \mu(X_t, t, \psi)dt + \sigma(X_t, t, \psi)dW_t \\ Y_t = f(X_t, \varepsilon_t), \quad \varepsilon_t \sim \pi(\varepsilon_t|\sigma_\varepsilon). \end{cases} \tag{1}$$

Define $\theta = (\psi, \sigma_\varepsilon)$ as the vector of unknowns to be estimated (might contain the initial state X_{t_0} as well). Data are discrete realisations $y_0, ..., y_n$ of $\{Y_t\}$.

Methods

- We consider the "early–rejection" ABC-MCMC approach described in [1]: basically **when using a uniform 0/1 kernel** for summary statistics comparison, *it is sometimes possible to avoid simulating from the model!* In some cases it reduces the computational time by 50-60%.
- Summary statistics are obtained using regression approaches, as in [2]: implemented methods are `mars` (multivariate adaptive regression splines) and lasso- type regularisation.
- Same as in [3], ABC tolerance δ is not chosen a-priori: a Markov chain is created for δ and parameters draws are selected ex-post among those corresponding to a "small enough" δ^*.

Early–rejection ABC-MCMC, see [1]

1. Initialization: choose or simulate $\theta_{start} \sim \pi(\theta)$, simulate $x_{start} \sim \pi(x|\theta_{start})$ and $y_{start} \sim \pi(y|x_{start}, \theta_{start})$. Fix $\delta_{start} > 0$ and $r = 0$. Starting values are $(\theta_r, \delta_r) \equiv (\theta_{start}, \delta_{start})$ and $S(y_{sim,r}) \equiv S(y_{start})$ such that $K(|S(y_{start}) - S(y)|/\delta_{start}) \equiv 1$. Here $K(\cdot)$ is the uniform 0/1 kernel.

At $(r+1)$th MCMC iteration:

2. generate $(\theta', \delta') \sim u(\theta, \delta|\theta_r, \delta_r)$ from its proposal distribution;
3. generate $\omega \sim U(0,1)$;

if
$$\omega > \frac{\pi(\theta')\pi(\delta')u(\theta_r, \delta_r|\theta', \delta')}{\pi(\theta_r)\pi(\delta_r)u(\theta', \delta'|\theta_r, \delta_r)} \quad (= \text{"ratio"})$$
then
$(\theta_{r+1}, \delta_{r+1}, S(y_{sim,r+1})) := (\theta_r, \delta_r, S(y_{sim,r}))$; ▷ (proposal rejected without simulating from the model)
else generate $x' \sim \pi(x|\theta')$ conditionally on the θ' from step 2; generate $y_{sim} \sim \pi(y|x', \theta')$ and calculate $S(y_{sim})$;
  if $K(|S(y_{sim}) - S(y)|/\delta') = 0$ **then**
    $(\theta_{r+1}, \delta_{r+1}, S(y_{sim,r+1})) := (\theta_r, \delta_r, S(y_{sim,r}))$ ▷ (proposal rejected)
  else if $\omega \leq$ ratio **then**
    $(\theta_{r+1}, \delta_{r+1}, S(y_{sim,r+1})) := (\theta', \delta', S(y_{sim}))$ ▷ (proposal accepted)
  else
    $(\theta_{r+1}, \delta_{r+1}, S(y_{sim,r+1})) := (\theta_r, \delta_r, S(y_{sim,r}))$ ▷ (proposal rejected)
  end if
end if

4. increment r to $r+1$ and go to step 2.

Main functions

- `abc_training`: performs a "'pilot" study to identify a region of the parameters space on which the expected value of the approximate posterior is likely to be, given a large number of simulated pairs of parameters from an initial prior $\theta_0 \sim \pi_0(\theta)$ and synthetic data y_{sim} conditionally on θ_0. From such datasets regression is performed to estimate $E(\theta|y_{sim})$ over many simulated y_{sim}, then from such sampling distribution for $\hat{E}(\theta|y_{sim})$ a prior $\pi(\theta)$ is deduced. $\pi(\theta)$ is the prior which will actually be used in the ABC-MCMC.
- `abc_mcmc`: "early–rejection" ABC-MCMC [1] using the prior deduced from `abc_training`. Notice `abc_mcmc` also produces a chain for δ.
- `abc_posthoc`: a graphical "post-hoc" determination of a "reasonable" δ^* applied on the output of `abc_mcmc`, same as in [3].

Some features of `abc-sde`

- Easy handling of *fully* and *partially observed* SDE systems;
- if no exact solution for the SDE in (1) is available, Euler-Maruyama integration is automatically performed;
- adaptive MCMC (Haario et al. 2001) is used to advance the simulation;
- uses `mars` and `lasso` for summary statistics determination;
- two case studies are provided in the `abc-sde` Reference Manual.

A toy model: stochastic Lotka-Volterra

Two chemical "species" interact via some reactions (not reported here) and the sizes $X_{t,1}$ and $X_{t,2}$ of the two "populations" at time t are simulated exactly using the "Gillespie algorithm". We add some Gaussian measurement error and obtain data $y_0, y_1, ..., y_{49}$, with $y_i \in \mathbb{R}^2_+$. We approximate the true underlying dynamics via a *chemical Langevin equation*:

$$dX_t = Sh(X_t, c)dt + \sqrt{S\text{diag}\{h(X_t, c)\}S^T}dW_t \tag{2}$$

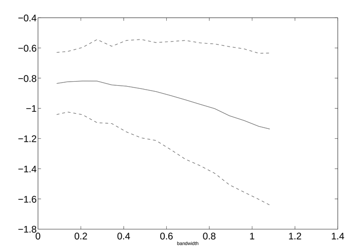
(c_1, c_2, c_3) are unknown constant-rates for the (not reported) chemical reactions, $h(X_t, c) = (c_1X_{t,1}, c_2X_{t,1}X_{t,2}, c_3X_{t,2})^T$ is the *hazard function* and S is the *stoichiometry matrix*:

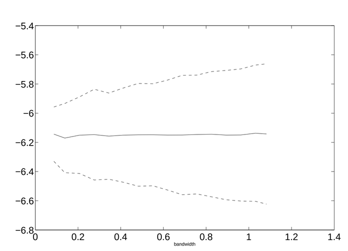
$$S = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{pmatrix}.$$

By using available data $y_0, y_1, ..., y_{49}$ incorporating Gaussian error with known variance σ^2_ε we wish to estimate the rates (c_1, c_2, c_3) (see [1] for a way more complex scenario).

Results 1

- The SDE is defined into `lv_sdefile.m`. We choose diffuse priors coded into `lv_prior.m` for the ABC "pilot": $\log c_1 \sim U(-3, 2)$, $\log c_2 \sim U(-7, 0)$, $\log c_3 \sim U(-3, 2)$. From the pilot results we deduce the following priors for the actual ABC-MCMC: $\log c_1 \sim N(-1.55, 0.4^2)$, $\log c_2 \sim N(-6, 0.3^2)$, $\log c_3 \sim N(-1.45, 0.5^2)$.
- the ABC-MCMC runs for two million iterations. We use `abc_posthoc()` to study the posterior means variation with increasing δ:


(a) $\log c_1$


(b) $\log c_2$

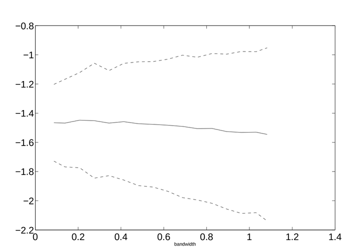

(c) $\log c_3$

Figure: LV model: posterior means for varying bandwidth δ.

- we deduce that we should keep draws corresponding to $\delta < 0.25$ as for larger bandwidths posterior means vary markedly (particularly for $\log c_1$).

Results 2

We *filter out* draws corresponding to $\delta > 0.25$ and use the remaining ones for posterior inference.
We obtain the following posterior means:

- c_1 : 0.44 [0.35,0.55], c_2 0.0021 [0.0017, 0.0028], c_3: 0.23 [0.17, 0.33]
- true values used to produce data are $(c_1^*, c_2^*, c_3^*) = (0.5, 0.0025, 0.3)$.
- a comparison between the true value of $\log c_1$, the kernel smoothing estimate of the ABC posterior density for $\log c_1$, its Gaussian prior used during the ABC-MCMC and the uniform prior used in the pilot:

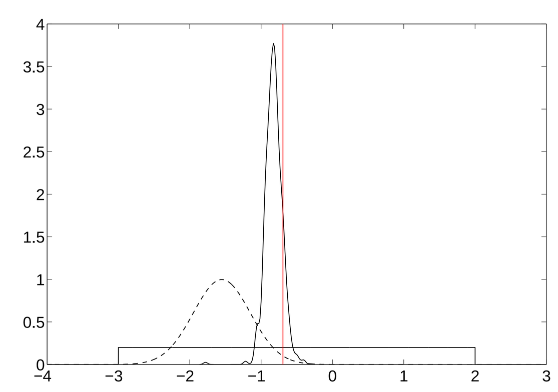


Figure: Approximate posterior for $\log c_1$ (solid curve), Gaussian prior (dashed line) and uniform prior used during the pilot. The vertical line corresponds to the true value of $\log c_1$.

Further possibilities

- It is straightforward to conduct inference for partially observed systems, where only one coordinate is observed ($X_{t,1}$ or $X_{t,2}$);
- it is also possible to estimate σ_ε as well as initial states $(X_{t_0,1}, X_{t_0,2})$.

See the `abc-sde` Reference Manual for further guidance. See [1] for a 4-dimensional SDE.

References

[1] Picchini, U. (2013). Inference for SDE models via approximate Bayesian computation. `arXiv:1204.5459`.

[2] Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. JRSS-B, 74(3), 419-474.

[3] Bortot, P., Coles, S.G., and Sisson, S.A. (2007). Inference for stereological extremes. JASA, 102(477), 84-92.