EasyABC: a R package to perform efficient approximate Bayesian computation sampling schemes

Franck Jabot, Thierry Faure, Nicolas Dumoulin

EasyABC version 1.2.99, 2014-02-25

Contents

[1] This document is included as a vignette (a LaTeX document created using the R function Sweave) of the

1 Summary

The aim of this vignette is to present the features of the `EasyABC` package. Section 2 describes the different algorithms available in the package. Section 3 details how to install the package and the formatting requirements. Sections 4 and 5 present two detailed worked examples.

2 Overview of the package EasyABC

`EasyABC` enables to launch various ABC schemes and to retrieve the ouputs of the simulations, so as to perform post-processing treatments with the various R tools available. `EasyABC` is also able to launch the simulations on multiple cores of a multi-core computer. Three main types of ABC schemes are available in EasyABC: the standard rejection algorithm of Pritchard et al. (1999), sequential schemes first proposed by Sisson et al. (2007), and coupled to MCMC schemes first proposed by Marjoram et al. (2003). Four different sequential algorithms are available: the ones of Beaumont et al. (2009), Drovandi and Pettitt (2011), Del Moral et al. (2012) and Lenormand et al. (2012). Three different MCMC schemes are available: the ones of Marjoram et al. (2003), Wegmann et al. (2009a) and a modification of Marjoram et al. (2003)'s algorithm in which the tolerance and proposal range are determined by the algorithm, following the modifications of Wegmann et al. (2009a). Details on how to implement these various algorithms with `EasyABC` are given in the manual pages of each function and two examples are detailed in Sections 4 and 5. We provide below a short presentation of each implemented algorithm.

2.1 The standard rejection algorithm of Pritchard et al. (1999)

This sampling scheme consists in drawing the model parameters in the prior distributions, in using these model parameter values to launch a simulation and in repeating this two-step procedure `nb_simul` times. At the end of the `nb_simul` simulations, the simulations closest to the target (or at a distance smaller than a tolerance threshold) in the space of the summary statistics are retained to form an approximate posterior distribution of the model parameters.

2.2 Sequential algorithms

Sequential algorithms for ABC have first been proposed by Sisson et al. (2007). These algorithms aim at reducing the required number of simulations to reach a given quality of the posterior approximation. The underlying idea of these algorithms is to spend more time in the areas of the parameter space where simulations are frequently close to the target. Sequential algorithms consist in a first step of standard rejection ABC, followed by a number of steps where the sampling of the parameter space is not anymore performed according to the prior distributions of parameter values. Various ways to perform this biased sampling have been proposed, and four of them are implemented in the package `EasyABC`.

2.3 Coupled to MCMC algorithms

The idea of ABC-MCMC algorithms proposed by Marjoram et al. (2003) is to perform a Metropolis-Hastings algorithm to explore the parameter space, and in replacing the step of likelihood ratio computation by simulations of the model. The original algorithm of Marjoram et al. (2003) is implemented in the method "Marjoram_original" in `EasyABC`. Wegmann et al. (2009) later proposed a number of improvements to the original scheme of Marjoram et al. (2003): they proposed to perform a calibration step so that the algorithm automatically determines the tolerance threshold, the scaling of the summary statistics and the scaling of the jumps in the parameter space during the MCMC. These improvements have been implemented in the method "Marjoram". Wegmann

package `EasyABC`. It is automatically dowloaded together with the package and can be accessed through `R` typing `vignette("EasyABC")`.

et al. (2009) also proposed additional modifications, among which a PLS transformation of the summary statistics. The complete Wegmann et al. (2009)'s algorithm is implemented in the method "Wegmann".

3 Installation and requirements

3.1 Installing the package

A version of R greater than or equal to 2.15.0 is required. The package has been tested on Windows 32 and Linux, but not on Mac. To install the `EasyABC` package from `R`, simply type:

```
> install.packages("EasyABC")
```

Once the package is installed, it needs to be loaded in the current `R` session to be used:

```
> library(EasyABC)
```

For online help on the package content, simply type:

```
> help(package="EasyABC")
```

For online help on a particular command (such as the function `ABC_sequential`), simply type:

```
> help(ABC_sequential)
```

3.2 The simulation code - for use on a single core

Users need to develop a simulation code with minimal compatibility constraints. The code can either be a `R` function or a binary executable file.

If the code is a `R` function, its argument must be a vector of parameter values and it must return a vector of summary statistics. If the option `use_seed=TRUE` is chosen, the first parameter value passed to the simulation code corresponds to the seed value to be used by the simulation code to initialize the pseudo-random number generator. The following parameters are the model parameters.

If the code is a binary executable file, it needs to read the parameter values in a file named 'input' in which each line contains one parameter value, and to output the summary statistics in a file named 'output' in which each summary statistics must be separated by a space or a tabulation. If the code is a binary executable file, a wrapper `R` function named 'binary_model' is available to interface the executable file with the `R` functions of the `EasyABC` package (see section 5 below).

Alternatively, users may prefer building a `R` function calling their binary executable file. A short tutorial is provided in section 3.8 to call a `C/C++` program.

3.3 The simulation code - for use with multiple cores

Users need to develop a simulation code with minimal compatibility constraints. The code can either be a `R` function or a binary executable file.

If the code is a `R` function, its argument must be a vector of parameter values and it must return a vector of summary statistics. The first parameter value passed to the simulation code corresponds to the seed value to be used by the simulation code to initialize the pseudo-random number generator. The following parameters are the model parameters. This means that the option `use_seed` must be turned to `TRUE` when using `EasyABC` with multiple cores.

If the code is a binary executable file, it needs to have as its single argument a positive integer `k`. It has to read the parameter values in a file named 'inputk' (where k is the integer passed as argument to the binary code: 'input1', 'input2'...) in which each line contains one parameter value, and to output the summary statistics in a file named 'outputk' (where k is the integer passed as argument to the binary code: 'output1', 'output2'...) in which each summary

statistics must be separated by a space or a tabulation. This construction avoids multiple cores to read/write in the same files. If the code is a binary executable file, a wrapper `R` function named 'binary_model_cluster' is available to interface the executable file with the `R` functions of the `EasyABC` package (see section 5 below).

Alternatively, users may prefer building a `R` function calling their binary executable file. A short tutorial is provided in section 3.8 to call a `C/C++` program.

3.4 Management of pseudo-random number generators

To insure that stochastic simulations are independent, the simulation code must either possess an internal way of initializing the seeds of its pseudo-random number generators each time the simulation code is launched. This can be achieved for instance by initializing the seed to the clock value. It is often desirable though to have a way to re-run some analyses with similar seed values. If this option is chosen, a seed value is provided in the input file as a first (additional) parameter, and incremented by 1 at each call of the simulation code. This means that the simulation code must be designed so that the first parameter is a seed initializing value. In the worked example (Section 5), the simulation code `trait_model` makes use of this package option, and in the first example (Section 4), the way this option can be used with a simple `R` function is demonstrated.

NB: Note that when using multicores with the package functions (`n_cluster=x` with `x` larger than 1), the option `use_seed=TRUE` is forced, since the seed value is also used to distribute the tasks to each core.

3.5 Encoding the prior distributions

A list encoding the prior distributions used for each model parameter must be supplied by the user. Each element of the list corresponds to a model parameter and can be defined by two ways:

1. By using predefined prior definition. The list element must be a vector whose first argument determines the type of prior distribution followed by the argument of the distribution function, possible values are:

   - "unif" for a uniform distribution on a segment, followed by two numbers the minimum and maximum values of the uniform distribution
   - "normal" for a normal distribution, followed by two numbers the mean and standard deviation of the normal distribution
   - "lognormal" for a lognormal distribution, followed by two numbers: the mean and standard deviation on the log scale of the lognormal distribution
   - "exponential" for an exponential distribution, followed by one number: the rate of the exponential distribution

   ```
   > my_prior=list(c("unif",0,1),c("normal",1,2))
   ```

2. By providing the sampling function and the density function. It means that you're free to provide your own methods. The providen data must be a list of two elements : the sampling function and the density function. For example, you can define a uniform distribution like that (equivalent to `my_prior=list(c("unif",0,1))`):

   ```
   > my_prior=list(list(c("runif",1,0,1), c("dunif",0,1)))
   ```

   Note that we should specify the first argument of "runif" to "1" (number of sampling). You can now add your custom method, like that:

   ```
   > my_prior=list(list(c("runif",1,0,1), c("dunif",0,1)),
   + list(c("mysample",arg1), c("mydensity", arg2, arg3)))
   ```

This scheme can be used to define discrete prior function. Let consider that you have a model parameter with n modalities from "1" to "n". You can define your uniform prior as follow:

```
> nbModalities=3
> mysample = function() { min(which(runif(1,0,nbModalities)<(1:nbModalities))) }
> mydensity = function(x) { 1/nbModalities }
> my_prior=list(list(c("mysample"), c("mydensity")))
```

3.6 The target summary statistics

A vector containing the summary statistics of the data must be supplied. The statistics must be in the same order as in the simulation outputs.

3.7 The option verbose

Intermediary results can be written in output files in the working directory. Users solely need to choose the option `verbose=TRUE` when launching the `EasyABC` functions (otherwise, the default value for `verbose` is `FALSE`). Intermediary results consist in the progressive writing of simulation outputs for the functions `ABC_rejection` and `ABC_mcmc` and in the writing of intermediary results at the end of each step for the function `ABC_sequential`. Additional details are provided in the help files of the functions.

3.8 Building a `R` function calling a `C/C++` program

Users having a `C/C++` simulation code may wish to construct a `R` function calling their `C/C++` program, instead of using the provided wrappers (see sections 3.2 and 3.3). The procedure is abundantly described in the 'Writing R Extensions' manual. In short, this can be done by:

- Adapt your C/C++ program by wrapping your main method into a `extern "C" { ... }` block. Here is an excerpt of the source code of the trait model provided in this package, in the folder `src`:

```
extern "C" {
  void trait_model(double *input,double *stat_to_return){
    // compute output and fill the array stat_to_return
  }
}
```

- Build your code into a binary library (.so under Linux or .dll under Windows) with the `R CMD SHLIB` command. In our example, the command for compiling the trait model and the given output are:

```
$ R CMD SHLIB trait_model_rc.cpp
g++ -I/usr/share/R/include -DNDEBUG -fpic -O2 -pipe -g -c trait_model_rc.cpp
  -o trait_model_rc.o
g++ -shared -o trait_model_rc.so trait_model_rc.o -L/usr/lib/R/lib -lR
```

- Load the builded library in your session with the `dyn.load` function.

```
> dyn.load("trait_model_rc.so")
```

- Use the `.C` function for calling your program, like we've done in our `trait_model` function:

```
trait_model <- function(input=c(1,1,1,1,1,1)) {
  .C("trait_model",input=c(input[1], 500, input[2:3], 1, input[4:5]),
    stat_to_return=array(0,4))$stat_to_return
}
```

You can also notice how we have fixed the parameter of the model.

3.9 Example of integration of an external program: `fastsimcoal`

This example is provided by an EasyABC user Albert Min-Shan Ko (currently at Department of genetics, Max Planck Institute of Evolutionary Anthropology, Leipzig, Germany). The purpose is to plug a third-party software related to population genetics into the EasyaABC workflow. This software needs input data in a given format, so the idea is to wrap the call to the `fastsimcoal` software into a script that will make the glue between EasyABC and `fastsimcoal`.

Here are the scripts as provided by courtesy of Albert Min-Shan Ko.

- The first one is a R script that will rewrite the sampled data into the format of a paramater for `fastsimcoal`.

```
r<-read.table('input',head=F)
sink('mod.input')
cat(paste('1','p1','unif',round(r[1,],0),round(r[1,],0),sep='\t'))
cat('\n')
cat(paste('1','p2','unif',round(r[2,],0),round(r[2,],0),sep='\t'))
cat('\n')
cat(paste('1','p3','unif',round(r[3,],0),round(r[3,],0),sep='\t'))
sink()
```

- Then a GNU Bash script that invoke this R script and integrate the well-formated parameters into a complete settings file, and run the program with this generated file (`sim.est`).

```
#!/bin/bash
rm -fr sim
Rscript mod.input.r
cat <(sed -n 1p template.est) <(sed -n '1,3'p mod.input) \
  <(sed -n '5,\$'p template.est) > sim.est
  until [ -f sim/arl_output ]; do
    ./fastsimcoal -t sim.tpl -e sim.est -E1 -n1 -q
    ./arlsumstat sim/sim_1_1.arp sim/arl_output 1 0 run_silent
  done
cat sim/arl_output > output
```

Then, the user can invoke EasyABC like this :

```
prior=list(c("unif",500,1000),c("unif",100,500),c("unif",50,200))
ABC_sim<-ABC_rejection(model=binary_model('./run_sim.sh'),prior=prior,nb_simul=3)
```

3.10 Example of integration of an java model

If your model runs with a Java Virtual Machine (can be written in Java, Scala, Groovy, ...), you can of course use the `binary_model` wrapper to run the JVM within your model. But, you can achieve a tigher integration that will simplify the process and save computing time. This section propose to use the R package `rJava`.

Let's consider the toy model written in Java (in a file named Model.java):

```
public class Model {
  public static double[] run(double[] x) {
      double[] result = new double[2];
      result[0] = x[0] + x[1];
      result[1] = x[0] * x[1];
      return result;
  }
}
```

We can compile it with the command: `javac Model.java` and then define our wrapper in R:

```
mymodel <- function(x) {
  library("rJava")
  .jinit(classpath=".")
  result = .jcall(J("Model"),"[D","run",.jarray(x))
  result
}
```

Then, the user can invoke EasyABC like this :

```
prior=list(c("unif",0,1),c("normal",1,2))
ABC_sim<-ABC_rejection(model=mymodel,prior=prior,nb_simul=3)
```

4 A first worked example

4.1 The toy model

We here consider a very simple stochastic model coded in the R language:

```
> toy_model<-function(x){
+         c( x[1] + x[2] + rnorm(1,0,0.1) , x[1] * x[2] + rnorm(1,0,0.1) )
+ }
```

We will use two different types of prior distribution for the two model parameters ($x[1]$ and $x[2]$): a uniform distribution between 0 and 1 and a normal distribution with mean 1 and standard deviation 2.

```
> toy_prior=list(c("unif",0,1),c("normal",1,2))
```

And we will consider an imaginary dataset of two summary statistics that the toy_model is aiming at fitting:

```
> sum_stat_obs=c(1.5,0.5)
```

4.2 Performing a standard ABC-rejection procedure

A standard ABC-rejection procedure can be simply performed with the function `ABC_rejection`, in precising the number n of simulations to be performed and the proportion of simulations which are to be retained p:

```
> set.seed(1)
> n=10
> p=0.2
> ABC_rej<-ABC_rejection(model=toy_model, prior=toy_prior, nb_simul=n,
+ summary_stat_target=sum_stat_obs, tol=p)
```

Alternatively, `ABC_rejection` can be used to solely launch the simulations and to store the simulation outputs without performing the rejection step. This option enables the user to make use of the `R` package `abc` (Csilléry et al. 2012) which offers an array of more sophisticated post-processing treatments than the simple rejection procedure:

```
> # Run the ABC rejection on the model
> set.seed(1)
> n=10
> ABC_rej<-ABC_rejection(model=toy_model, prior=toy_prior, nb_simul=n)

> # Install if needed the "abc" package
> install.packages("abc")

> # Post-process the simulations outputs
> library(abc)
> rej<-abc(sum_stat_obs, ABC_rej$param, ABC_rej$stats, tol=0.2, method="rejection")
> # simulations selected:
> rej$unadj.values
> # their associated summary statistics:
> rej$ss
> # their normalized euclidean distance to the data summary statistics:
> rej$dist
```

4.3 Performing a sequential ABC scheme

Other functions of the `EasyABC` package are used in a very similar manner. To perform the algorithm of Beaumont et al. (2009), one needs to specify the sequence of tolerance levels *tolerance_tab* and the number *nb_simul* of simulations to obtain below the tolerance level at each iteration:

```
> n=10
> tolerance=c(1.25,0.75)
> ABC_Beaumont<-ABC_sequential(method="Beaumont", model=toy_model,
+ prior=toy_prior, nb_simul=n, summary_stat_target=sum_stat_obs,
+ tolerance_tab=tolerance)
```

To perform the algorithm of Drovandi and Pettitt (2011), one needs to specify four arguments: the initial number of simulations *nb_simul*, the final tolerance level *tolerance_tab*, the proportion α of best-fit simulations to update the tolerance level at each step, and the target proportion c of unmoved particles during the MCMC jump. Note that default values $alpha = 0.5$ and $c = 0.01$ are used if not specified, following Drovandi and Pettitt (2011).

```
> n=10
> tolerance=0.75
> c_drov=0.7
> ABC_Drovandi<-ABC_sequential(method="Drovandi", model=toy_model,
+ prior=toy_prior, nb_simul=n, summary_stat_target=sum_stat_obs,
+ tolerance_tab=tolerance, c=c_drov)
```

To perform the algorithm of Del Moral et al. (2012), one needs to specify five arguments: the initial number of simulations *nb_simul*, the number α controlling the decrease in effective sample size of the particle set at each step, the number M of simulations performed for each particle, the minimal effective sample size *nb_threshold* below which a resampling of particles is performed and the final tolerance level *tolerance_target*. Note that default values $alpha = 0.5$, $M = 1$ and $nb\_threshold = nb\_simul/2$ are used if not specified.

```
> n=10
> alpha_delmo=0.5
```

```
> tolerance=0.75
> ABC_Delmoral<-ABC_sequential(method="Delmoral", model=toy_model,
+ prior=toy_prior, nb_simul=n, summary_stat_target=sum_stat_obs,
+ alpha=alpha_delmo, tolerance_target=tolerance)
```

To perform the algorithm of Lenormand et al. (2012), one needs to specify three arguments: the initial number of simulations *nb_simul*, the proportion α of best-fit simulations to update the tolerance level at each step, and the stopping criterion *p_acc_min*. Note that default values $alpha = 0.5$ and $p\_acc\_min = 0.05$ are used if not specified, following Lenormand et al. (2012). Also note that the method "Lenormand" is only supported with uniform prior distributions (since it performs a Latin Hypercube sampling at the beginning). Here, we therefore need to alter the prior distribution of the second model parameter:

```
> toy_prior2=list(c("unif",0,1),c("unif",0.5,1.5))
```

```
> n=10
> pacc=0.4
> ABC_Lenormand<-ABC_sequential(method="Lenormand", model=toy_model,
+ prior=toy_prior2, nb_simul=10, summary_stat_target=sum_stat_obs,
+ p_acc_min=pacc)
```

4.4 Performing a ABC-MCMC scheme

To perform the algorithm of Marjoram et al. (2003), one needs to specify five arguments: the number of sampled points *n_rec* in the Markov Chain, the number of chain points between two sampled points *n_between_sampling*, the maximal distance accepted between simulations and data *dist_max*, a vector *tab_normalization* precising the scale of each summary statistics, and a vector *proposal_range* precising the maximal distances in each dimension of the parameter space for a jump of the MCMC. All these arguments have default values (see the package help for the function `ABC_mcmc`), so that `ABC_mcmc` will work without user-defined values.

```
> n=10
> ABC_Marjoram_original<-ABC_mcmc(method="Marjoram_original", model=toy_model,
+ prior=toy_prior, summary_stat_target=sum_stat_obs, n_rec=n)
```

To perform the algorithm of Marjoram et al. (2003) in which some of the arguments (*dist_max*, *tab_normalization* and *proposal_range*) are automatically determined by the algorithm via an initial calibration step, one needs to specify three arguments: the number *n_calibration* of simulations to perform at the calibration step, the tolerance quantile *tolerance_quantile* to be used for the determination of *dist_max* and the scale factor *proposal_phi* to determine the proposal range. These modifications are drawn from the algorithm of Wegmann et al. (2009a), without relying on PLS regressions. The arguments are set by default to: $n\_calibration = 10000$, $tolerance\_quantile = 0.01$ and $proposal\_phi = 1$. This way of automatic determination of *dist_max*, *tab_normalization* and *proposal_range* is strongly recommended, compared to the crude automatic determination proposed in the method `Marjoram_original`.

```
> n=10
> ABC_Marjoram<-ABC_mcmc(method="Marjoram", model=toy_model,
+ prior=toy_prior, summary_stat_target=sum_stat_obs, n_rec=n)
```

To perform the algorithm of Wegmann et al. (2009a), one needs to specify four arguments: the number *n_calibration* of simulations to perform at the calibration step, the tolerance quantile *tolerance_quantile* to be used for the determination of *dist_max*, the scale factor *proposal_phi* to determine the proposal range and the number of components *numcomp* to be used in PLS regressions. The arguments are set by default to: $n\_calibration = 10000$, $tolerance\_quantile = 0.01$, $proposal\_phi = 1$ and $numcomp = 0$, this last default value encodes a choice of a number of PLS components equal to the number of summary statistics.

```
> n=10
> ABC_Wegmann<-ABC_mcmc(method="Wegmann", model=toy_model,
+ prior=toy_prior, summary_stat_target=sum_stat_obs, n_rec=n)
```

4.5 Using multiple cores

The functions of the package `EasyABC` can launch the simulations on multiple cores of a computer: users have to indicate the number of cores they wish to use in the argument `n_cluster` of the functions, and they have to use the option `use_seed=TRUE`. Users also need to design their code in a slightly different way so that it is compatible with the option `use_seed=TRUE` (see Section 3.3 for additional details). For the toy model above, the modifications needed are the following:

```
> toy_model_parallel<-function(x){
+         set.seed(x[1]) # so that each core is initialized with a different seed value.
+         c( x[2] + x[3] + rnorm(1,0,0.1) , x[2] * x[3] + rnorm(1,0,0.1) )
+ }
```

```
> set.seed(1)
> n=10
> p=0.2
> ABC_rej<-ABC_rejection(model=toy_model_parallel, prior=toy_prior,
+ nb_simul=n, summary_stat_target=sum_stat_obs, tol=p, n_cluster=2,
+ use_seed=TRUE)
```

5 A second worked example

5.1 The trait model

We turn now to a stochastic ecological model hereafter called `trait_model` to illustrate how to use `EasyABC` with models not initially coded in the `R` language. `trait_model` represents the stochastic dynamics of an ecological community where each species is represented by a set of traits (i.e. characteristics) which determine its competitive ability. A detailed description and analysis of the model can be found in Jabot (2010). The model requires four parameters: an immigration rate I, and three additional parameters (h, A and σ) describing the way traits determine species competitive ability. The model additionnally requires two fixed variables: the total number of individuals in the local community J and the number of traits used n_t. The model outputs four summary statistics: the species richness of the community S, its Shannon's index H, the mean of the trait value among individuals MTV and the skewness of the trait value distribution STV.

NB: Three parameters (I, A and σ) have non-uniform prior distributions: instead, their log-transformed values have a uniform prior distribution. The simulation code `trait_model` therefore takes an exponential transform of the values proposed by `EasyABC` for these parameters at the beginning of each simulation.

In the following, we will use the values $J = 500$ and $n\_t = 1$, and uniform prior distributions for $ln(I)$ in [3; 5], h in [-25;125], $ln(A)$ in $[ln(0.1); ln(5)]$ and $ln(\sigma)$ in $[ln(0.5); ln(25)]$. The simulation code `trait_model` reads sequentially J, I, A, n_t, h and σ.

NB: Note that the fixed variables J and n_t have been fixed (see section RClink) into the function `trait_model` (in versions previous to 1.3, they was included in the prior list using uniform distributions with a trivial ranges)

```
> trait_prior=list(c("unif",3,5),c("unif",-2.3,1.6),
+ c("unif",-25,125), c("unif",-0.7,3.2))
```

We will consider an imaginary dataset whose summary statistics are $(S, H, MTV, STV) = (100, 2.5, 20, 30000)$:

```
> sum_stat_obs=c(100,2.5,20,30000)
```

5.2 Performing a standard ABC-rejection procedure

A standard ABC-rejection procedure can be simply performed with the function `ABC_rejection`, in precising the number n of simulations to be performed and the proportion p of retained simulations. Note that the option `use_seed=TRUE` is used, since `trait_model` requires a seed initializing value for its pseudo-random number generator:

```
> set.seed(1)
> n=10
> p=0.2
> ABC_rej<-ABC_rejection(model=trait_model, prior=trait_prior, nb_simul=n,
+ summary_stat_target=sum_stat_obs, tol=p, use_seed=TRUE)
```

Alternatively, `ABC_rejection` can be used to solely launch the simulations and to store the simulation outputs without performing the rejection step. This option enables the user to make use of the `R` package `abc` (Csilléry et al. 2012) which offers an array of more sophisticated post-processing treatments than the simple rejection procedure:

```
> install.packages("abc")

> library(abc)
> set.seed(1)
> n=10
> p=0.2
> ABC_rej<-ABC_rejection(model=trait_model, prior=trait_prior, nb_simul=n, use_seed=TRUE)
> rej<-abc(sum_stat_obs, ABC_rej$param, ABC_rej$stats,
+ tol=0.2, method="rejection")
> # simulations selected:
> rej$unadj.values
> # their associated summary statistics:
> rej$ss
> # their normalized euclidean distance to the data summary statistics:
> rej$dist
```

Note that a simulation code `My_simulation_code` can be passed to the function `ABC_rejection` in several ways depending on its nature:

- if it is a `R` function
  `ABC_rejection(My_simulation_code, prior, nb_simul,...)`
- if it is a binary executable file and a single core is used (see section 3.2 for compatibility constraints)
  `ABC_rejection(binary_model("./My_simulation_code"), prior, nb_simul, use_seed=TRUE,...)`
- if it is a binary executable file and multiple cores are used (see section 3.3 for compatibility constraints)
  `ABC_rejection(binary_model_cluster("./My_simulation_code"), prior, nb_simul, n_cluster=2, use_seed=TRUE)`

5.3 Performing a sequential ABC scheme

Other functions of the `EasyABC` package are used in a very similar manner. To perform the algorithm of Beaumont et al. (2009), one needs to specify the sequence of tolerance levels *tolerance_tab* and the number *nb_simul* of simulations to obtain below the tolerance level at each iteration:

```
> n=10
> tolerance=c(8,5)
```

```
> ABC_Beaumont<-ABC_sequential(method="Beaumont", model=trait_model,
+ prior=trait_prior, nb_simul=n, summary_stat_target=sum_stat_obs,
+ tolerance_tab=tolerance, use_seed=TRUE)
```

To perform the algorithm of Drovandi and Pettitt (2011), one needs to specify four arguments: the initial number of simulations *nb_simul*, the final tolerance level *tolerance_tab*, the proportion α of best-fit simulations to update the tolerance level at each step, and the target proportion c of unmoved particles during the MCMC jump. Note that default values $alpha = 0.5$ and $c = 0.01$ are used if not specified, following Drovandi and Pettitt (2011).

```
> n=10
> tolerance=3
> c_drov=0.7
> ABC_Drovandi<-ABC_sequential(method="Drovandi", model=trait_model,
+ prior=trait_prior, nb_simul=n, summary_stat_target=sum_stat_obs,
+ tolerance_tab=tolerance, c=c_drov, use_seed=TRUE)
```

To perform the algorithm of Del Moral et al. (2012), one needs to specify five arguments: the initial number of simulations *nb_simul*, the number α controlling the decrease in effective sample size of the particle set at each step, the number M of simulations performed for each particle, the minimal effective sample size *nb_threshold* below which a resampling of particles is performed and the final tolerance level *tolerance_target*. Note that default values $alpha = 0.5$, $M = 1$ and $nb\_threshold = nb\_simul/2$ are used if not specified.

```
> n=10
> alpha_delmo=0.5
> tolerance=3
> ABC_Delmoral<-ABC_sequential(method="Delmoral", model=trait_model,
+ prior=trait_prior, nb_simul=n, summary_stat_target=sum_stat_obs,
+ alpha=alpha_delmo, tolerance_target=tolerance, use_seed=TRUE)
```

To perform the algorithm of Lenormand et al. (2012), one needs to specify three arguments: the initial number of simulations *nb_simul*, the proportion α of best-fit simulations to update the tolerance level at each step, and the stopping criterion *p_acc_min*. Note that default values $alpha = 0.5$ and $p\_acc\_min = 0.05$ are used if not specified, following Lenormand et al. (2012).

```
> n=10
> pacc=0.4
> ABC_Lenormand<-ABC_sequential(method="Lenormand", model=trait_model,
+ prior=trait_prior, nb_simul=n, summary_stat_target=sum_stat_obs,
+ p_acc_min=pacc, use_seed=TRUE)
```

5.4 Performing a ABC-MCMC scheme

To perform the algorithm of Marjoram et al. (2003), one needs to specify five arguments: the number of sampled points *n_obs* in the Markov Chain, the number of chain points between two sampled points *n_between_sampling*, the maximal distance accepted between simulations and data *dist_max*, a vector *tab_normalization* precising the scale of each summary statistics, and a vector *proposal_range* precising the maximal distances in each dimension of the parameter space for a jump of the MCMC. All these arguments have default values (see the package help for the function `ABC_mcmc`), so that `ABC_mcmc` will work without user-defined values.

```
> n=10
> ABC_Marjoram_original<-ABC_mcmc(method="Marjoram_original", model=trait_model,
+ prior=trait_prior, summary_stat_target=sum_stat_obs, n_rec=n, use_seed=TRUE)
```

To perform the algorithm of Marjoram et al. (2003) in which some of the arguments (*dist_max*, *tab_normalization* and *proposal_range*) are automatically determined by the algorithm via an initial calibration step, one needs to specify three arguments: the number *n_calibration* of simulations to perform at the calibration step, the tolerance quantile *tolerance_quantile* to be used for the determination of *dist_max* and the scale factor *proposal_phi* to determine the proposal range. These modifications are drawn from the algorithm of Wegmann et al. (2009a), without relying on PLS regressions. The arguments are set by default to: *n_calibration* = 10000, *tolerance_quantile* = 0.01 and *proposal_phi* = 1. This way of automatic determination of *dist_max*, *tab_normalization* and *proposal_range* is strongly recommended, compared to the crude automatic determination proposed in the method `Marjoram_original`.

```
> n=10
> n_calib=10
> tol_quant=0.2
> ABC_Marjoram<-ABC_mcmc(method="Marjoram", model=trait_model, prior=trait_prior,
+ summary_stat_target=sum_stat_obs,
+ n_rec=n, n_calibration=n_calib, tolerance_quantile=tol_quant, use_seed=TRUE)
```

To perform the algorithm of Wegmann et al. (2009a), one needs to specify four arguments: the number *n_calibration* of simulations to perform at the calibration step, the tolerance quantile *tolerance_quantile* to be used for the determination of *dist_max*, the scale factor *proposal_phi* to determine the proposal range and the number of components *numcomp* to be used in PLS regressions. The arguments are set by default to: *n_calibration* = 10000, *tolerance_quantile* = 0.01, *proposal_phi* = 1 and *numcomp* = 0, this last default value encodes a choice of a number of PLS components equal to the number of summary statistics.

```
> n=10
> n_calib=10
> tol_quant=0.2
> ABC_Wegmann<-ABC_mcmc(method="Wegmann", model=trait_model, prior=trait_prior,
+ summary_stat_target=sum_stat_obs,
+ n_rec=n, n_calibration=n_calib, tolerance_quantile=tol_quant, use_seed=TRUE)
```

5.5 Using multiple cores

The functions of the package `EasyABC` can launch the simulations on multiple cores of a computer: users only have to indicate the number of cores they wish to use in the argument `n_cluster` of the functions. The compatibility constraints of the simulation code are slightly different when using multiple cores: please refer to section 3.3 for more information.

6 Troubleshooting and development

Please send comments, suggestions and bug reports to nicolas.dumoulin@irstea.fr or franck.jabot@irstea.fr
Any new development of more efficient ABC schemes that could be included in the package is particularly welcome.

7 Programming Acknowledgements

The `EasyABC` package makes use of a number of `R` tools, among which:
- the `R` package `lhs` (Carnell 2012) for latin hypercube sampling.
- the `R` package `MASS` (Venables and Ripley 2002) for boxcox transformation.
- the `R` package `mnormt` (Genz and Azzalini 2012) for multivariate normal generation.
- the `R` package `pls` (Mevik and Wehrens 2011) for partial least square regression.
- the `R` script for the Wegmann et al. (2009a)'s algorithm drawn from the `ABCtoolbox` documentation (Wegmann et al. 2009b).

8 References

Beaumont, M. A., Cornuet, J., Marin, J., and Robert, C. P. (2009) Adaptive approximate Bayesian computation. *Biometrika*,**96**, 983–990.

Carnell, R. (2012) lhs: Latin Hypercube Samples. R package version 0.10. http://CRAN.R-project.org/package=lhs

Csilléry, K., François, O., and Blum, M.G.B. (2012) abc: an r package for approximate bayesian computation (abc). *Methods in Ecology and Evolution*, **3**, 475–479.

Del Moral, P., Doucet, A., and Jasra, A. (2012) An adaptive sequential Monte Carlo method for approximate Bayesian computation. *Statistics and Computing*, **22**, 1009–1020.

Drovandi, C. C. and Pettitt, A. N. (2011) Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. *Biometrics*, **67**, 225–233.

Genz, A., and Azzalini, A. (2012) mnormt: The multivariate normal and t distributions. R package version 1.4-5. http://CRAN.R-project.org/package=mnormt

Jabot, F. (2010) A stochastic dispersal-limited trait-based model of community dynamics. *Journal of Theoretical Biology*, **262**, 650–661.

Lenormand, M., Jabot, F., Deffuant G. (2012) Adaptive approximate Bayesian computation for complex models. http://arxiv.org/pdf/1111.1308.pdf

Marjoram, P., Molitor, J., Plagnol, V. and Tavaré, S. (2003) Markov chain Monte Carlo without likelihoods. *PNAS*, **100**, 15324–15328.

Mevik, B.-H., and Wehrens, R. (2011) pls: Partial Least Squares and Principal Component regression. R package version 2.3-0. http://CRAN.R-project.org/package=pls

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

Sisson, S.A., Fan, Y., and Tanaka, M.M. (2007) Sequential Monte Carlo without likelihoods. *PNAS*, **104**, 1760–1765.

Venables, W.N., and Ripley, B.D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York.

Wegmann, D., Leuenberger, C. and Excoffier, L. (2009a) Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics*, **182**, 1207-1218.

Wegmann, D., Leuenberger, C. and Excoffier, L. (2009b) Using ABCtoolbox. http://cmpg.unibe.ch/software/abctoolbox/ABCtoolbox_manual.pdf