

Manual for ABC-acoustic version: 1.0

S.J. Puechmaille

Date: 15/04/2018

Please, report any bugs or suggestions to improve the script, input or output files to: s [dot] puechmaille [at] gmail [dot] com

This script is intended to become an R package. Please, contact the author should you be interested in contributing to it.

Before reporting on a bug, please, make sure your input files are in the correct format.

TABLE OF CONTENTS

BEFORE RUNNING ANALYSES.....	1
ANALYSES	1
I –Running analyses on simulated datasets- Script 1.	2
A- Input files	2
B- Important arguments/parameters to set for the simulations.....	2
C- Storing the values/estimates in an object.....	2
D- Running the simulations and ABC (and Exclusion) estimates	3
II –Running analyses from empirical datasets - Script 2	4
A- Input files	4
B- Important arguments/parameters to prepare the empirical dataset	4
C- Important arguments/parameters for the analyses	5
D- Storing the values/estimates in an object.....	5
E- Running the simulations and ABC (and Exclusion) estimates	6
CHANGES FROM PREVIOUS VERSIONS	6
REFERENCES	6

Here we provide some information on how to estimate POM (the proportion of males) from acoustic data (or any other data with similar structure). The details about the method used are described in:

- Lenormand *et al.* (2013) for the ABC algorithm used
- Lehnen *et al.* (submitted) for the ABC and exclusion approaches,

BEFORE RUNNING ANALYSES

You first need to make sure you have the necessary libraries installed. The necessary libraries are 'EasyABC' and 'HDInterval' and 'reshape2'. If you do not have these libraries installed, simply type the following commands in R:

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

then select a mirror of choice and wait until the first library is installed. Then install the second library by typing:

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

followed by the third library:

```
>install.packages("reshape2")
```

and the fourth library:

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

Once you have the libraries installed, you have to load them by typing:

```
>library(EasyABC)
```

```
>library(HDInterval)
```

```
>library(reshape2)
```

```
>library(plyr)
```

Before you run any analyses, it is best to locate where R is working so that you know where to place your input file (with acoustic data) and where R will save your output results. To find out the R working directory, simply type in the console:

```
>getwd()
```

If R is working where you want, proceed with the next step. If not, you can change the working directory using the 'setwd' command. For example, to work in the following directory "D:/SEB/R/R-3.4.0patched/adata/Lisa", simply type the following command:

```
>setwd("D:/SEB/R/R-3.4.0patched/adata/Lisa")
```

Please, note that Windows is using backslash ('\\') for defining the path while R is using (forward) slash ('/').

ANALYSES

The scripts used in Lehnen *et al.* (2018) are provided. These are adapted to run the analyses as presented in the paper. The first script, detailed in paragraph I below, can be used to run analyses on simulated datasets (to test the performance of the method under certain conditions provided by the user). The second script, detailed in paragraph II below, can be used to run analyses on empirical datasets.

I –Running analyses on simulated datasets- Script 1.

A- Input files

No input file is needed here as this script is used for testing the methods based on simulated datasets that are created as part of the script.

B- Important arguments/parameters to set for the simulations

In Part II, step 1 & 2, the user can set the parameters and priors of the simulations and analyses. Here is a brief description of these parameters.

Step 1

Pmale= Vector or the proportion of males considered (values have to be between 0 and 1)

Ntot= Vector of the number of calls considered for the simulations

alpha_Lenormand= Vector of values for the Alpha parameter of the ABC algorithm

pacc= Vector of values for the pacc parameter of the 'ABC_sequential' function.

For the 4 arguments above, you can specify either single or multiple values. If multiple values of some arguments are provided, the script will calculate all possible combinations of the arguments (as provided by the 'expand.grid' function) and run simulations accordingly. The combination of parameters is available in the 'mordec' object.

Nrep= Number of simulated datasets.

nb_simul = Initial number of simulations within the ABC approach (= nb_simul parameter of the 'ABC_sequential' function).

MaleFreq= mean value for the males acoustic parameter (here mean peak frequency)

FemaleFreq= mean value for the females acoustic parameter (here mean peak frequency)

SdFreqFM= Standard deviation of the acoustic parameter (here mean peak frequency); this value is considered as being the same for males and females.

Run= Run number (if results are exported, the folder containing them will have this name)

Step 2

Rh_prior= list of prior information for the ABC algorithm. Each element of the list corresponds to a model parameter. The list element must be a vector whose first argument determines the type of prior distribution: the only possible value is "unif" for a uniform distribution when using the method "Lenormand". The following arguments of the list elements contain the characteristics of the prior distribution chosen: for "unif", two numbers must be given: the minimum and maximum values of the uniform distribution.

Rh_Constraints= a string expressing the constraints between model parameters. This expression will be evaluated as a logical expression, you can use all the logical operators including "<"< code=" ">, ">", etc. Each parameter should be designated with "X1", "X2", etc. in the same order as in the prior definition.

C- Storing the values/estimates in an object

Although this is automatically done and does not require your attention, the results from each ABC (and Exclusion) analysis will be stored in an object called ResA. This allows

large simulations to be carried out (with multiple parameters combinations; see Step 1 above) while the results are saved for latter comparison or further analyses. ResA is an array with the following structure.

Dimension 1 (=lines): Results from the Nrep simulated datasets.

Dimension 2 (=columns): Combines estimated values, information about the data, simulation parameters, errors, etc. as detailed below:

```
#For each parameters (pM: proportion of males, fM: average peak frequency of males,
  fF: average peak frequency of females, SD: standard deviation of peak frequency)
  #the mean (mea), median (Med), lower (hl) and upper (hh) bound of the 95% highest
  density interval of the estimate are presented.
  #e.g. pMmea is the mean proportion of males as estimated by the ABC approach.
  #e.g. fFhh is the upper bound of the 95% highest density interval of the estimate for
  the average peak frequency for females
#For each of the exclusion method thresholds (95, 99 and 99.9% CI), 5 columns are
  presented
  #Value of the peak frequency threshold above which 95% of females echolocate at
  (e.g. sFl95)
  #Number of females that have a peak frequency higher than the value of the peak
  frequency threshold above which 95% of females echolocate at (e.g. nF95)
  #Value of the peak frequency threshold under which 95% of males echolocate at
  (e.g. sMh95)
  #Number of males that have a peak frequency lower than the value of the peak
  frequency threshold under which 95% of males echolocate at (e.g. nM95)
  #Proportion of males (e.g. pMp95) calculated from the above mentioned number of
  males (e.g. nM95) and females (e.g. nF95)
#MeaRMSE: ABC estimated minus true value of the proportion of males (used for
  RMSE calculation)
#pMp95RMSE, pMp99RMSE, pMp999RMSE: exclusion method estimate (95, 99 and
  99.9%CI respectively) of the proportion of males minus true value (used for RMSE
  calculation)
#Pacc: pacc parameter of the ABC algorithm
#Alpha_Len: Alpha parameter of the ABC algorithm
#Nbsim: Initial number of simulations = nb_simul
#TRUEpM: true proportion of males (only applicable when running simulations)
#Ntot: number of individual points in the dataset
#Nmal: number of male points in the dataset
#Nfem: number of female points in the dataset
```

Dimension 3 (=matrices): Results from each combination of parameters as defined in the 'mordec' object are presented in a different matrix.

D- Running the simulations and ABC (and Exclusion) estimates

No input is required from the user in this Part III. Outputs of this section can be found in a folder named according to the 'Run' parameter (cf. I-B above); the folder is located in the working directory (simply type in the console 'getwd()' if you do not know where your working directory is). Two files are exported from R:

1-A '.csv' file with the results present in the 'ResA' array. The file is named according to the date of the run; e.g. 'ABC_2018-01-30_Results.csv' for a run carried out on 30.01.2018. The column X represents the first dimension of the array, Y the second dimension and Z the third

dimension. The associated values are presented in the 'value' column. This .csv file can be re-imported in R and reformatted into an array using the following commands:

```
>EpiC<-read.csv("D:/SEB/R/R-3.4.0patched/adata/Run-Test1/ ABC_2018-01-30_Results.csv",h=TRUE)
>Epiq<-acast(EpiC, X~ Z ~ Y)
```

II –Running analyses from empirical datasets - Script 2

The backbone of this script is similar to script 1.

A- Input files

One input file with the empirical dataset is needed here. The file contains data in 3 columns (with the headers as presented below) !!!!! [without these exact headers, the script will not work] !!!!!

#Column 1: named 'Fmean': peak frequency (or any acoustic parameter of interest)

#Column 2: named 'Colo': name of the Colony (or site or any other level of interest)

#Column 3: named 'Filename': File name (indicating the grouping or the calls).

The content of the file should start as follows:

Fmean,Colo,Filename

108.52,Thu22,1

108.11,Thu22,1

108.37,Thu22,1

108.11,Thu22,1

107.66,Thu22,1

108.52,Thu22,1

108.37,Thu22,1

109.59,Thu22,2

109.89,Thu22,2

111.11,Thu22,2

An example input file can be found as supplementary information with the Lehnen et al. (2018, PLoS ONE) paper. Please, note that in some countries (e.g. France, Germany) the coma is used as a decimal separator and this sign is therefore not available to be used as a separator in csv files. Instead, the semi-colon “;” is used. In case you encounter problems, either set your computer to use the international standard for decimals or modify the argument ‘sep’ in the lines with ‘read.csv’ to be able to correctly import .csv files.

B- Important arguments/parameters to prepare the empirical dataset

Part I must be carefully checked by the user as there are multiple occasions where user input is needed. Here is a brief description of these arguments/parameters.

Colony= Colony of interest (as per the information appearing in the 'Colo' column)

PF.Thresh= Threshold of peak frequency under which the calls are removed from the dataset

Level= select if you want your data to be averaged over 'Recordings' or if you want to carry out the analysis at the 'Calls' level

extra= Extra info to use in the file name (for later exporting of results). This information plus other information provided by the user is then used automatically to generate a Run 'name' that will be used if files/results are exported.

FilterSD= To filter out Recordings with high standard deviation, set the 'FilterSD' parameter to 'TRUE' , otherwise, to 'FALSE'. Note that this option should only be used if "Level" is set to "Recordings".

SD.Thresh = Threshold of sd above which the recordings are removed from the dataset. Note that this value must be set when "FilterSD" is set to TRUE.

C- Important arguments/parameters for the analyses

In the Part II, step 1 & 2, the user can set the parameters and priors of the analyses. Here is a brief description of these parameters.

Step 1

Nrep= Number of repetitions (if needed, the ABC analyses can be repeated multiple times for one empirical dataset).

nb_simul = Initial number of simulations within the ABC approach (= nb_simul parameter of the 'ABC_sequential' function).

alpha_Lenormand= Vector of values for the Alpha parameter of the ABC algorithm (=alpha_Lenormand parameter of the 'ABC_sequential' function)

pacc= Vector of values for the pacc parameter of the 'ABC_sequential' function.

For the 2 arguments above, you can specify either single or multiple values. If multiple values of some arguments are provided, the script will calculate all possible combinations of the arguments (as provided by the 'expand.grid' function) and run simulations accordingly. The combination of parameters are available in the 'mordec' object.

Step 2

Rh_prior= list of prior information for the ABC algorithm. Each element of the list corresponds to a model parameter. The list element must be a vector whose first argument determines the type of prior distribution: the only possible value is "unif" for a uniform distribution when using the method "Lenormand". The following arguments of the list elements contain the characteristics of the prior distribution chosen. For "unif", two numbers must be given: the minimum and maximum values of the uniform distribution.

Rh_Constraints= a string expressing the constraints between model parameters. This expression will be evaluated as a logical expression, you can use all the logical operators including "<"< code="">, ">", etc. Each parameter should be designated with "X1", "X2", etc. in the same order as in the prior definition.

D- Storing the values/estimates in an object

Although this is automatically done and does not require your attention, the results from

each ABC (and Exclusion) analysis will be stored in an object called ResA (Part II, Step 3 of the script). This object is exactly the same as detailed in the section I-C above. No further description is provided below. Please, note that some columns of the 'ResA' array do not have much meaning when analyzing empirical datasets (e.g. MeaRMSE) though these are kept here so that the array with results is identical between the analysis of empirical and simulated datasets.

E- Running the simulations and ABC (and Exclusion) estimates

No input is required from the user in this Part II, step 4. Outputs of this section can be found in a folder named according to the 'Run' value (cf. II-B above); the folder is located in the working directory (simply type in the console 'getwd()' if you do not know where your working directory is). Two files are exported from R:

1-A '.csv' file with the results present in the 'ResA' array. The file is named 'ABC_Results.csv'. The column X represents the first dimension of the array, Y the second dimension and Z the third dimension. The associated values are presented in the 'value' column. This .csv file can be re-imported in R and reformatted into an array using the following commands:

```
>EpiC<-read.csv("D:/SEB/R/R-3.4.0patched/adata/Run_Thu47_Recordings_Example/
ABC_Results.csv",h=TRUE)
>Epiq<-acast(EpiC, X~ Z ~ Y)
```

Please note that if R returns the error message below, there is nothing to worry about as the issue is with the p-values in the Kolmogorov-smirnov test but we do not use the p-values here, only the Kolmogorov-smirnov distances which are therefore not affected.

Warnings 'In ks.test(c(MAL, FEM), c(dataE\$mean), exact = FALSE) : p-value will be approximate in the presence of ties'

CHANGES FROM PREVIOUS VERSIONS

This is the first version of the script.

REFERENCES

- Lehnen L, Schorcht W, Karst I, Biedermann M, Kerth G, Puechmaille SJ (submitted) Using Approximate Bayesian Computation to infer sex ratios from acoustic data. PLoS ONE
- Lenormand M, Jabot F, Deffuant G (2013) Adaptive approximate Bayesian computation for complex models. Computational Statistics 28:2777-2796. doi:10.1007/s00180-013-0428-3