**User Manual of DzageEM**



Jin-Tao Kong

Jilin University

Index

[1. Introduction 3](#_Toc86386274)

[2. Data import format 4](#_Toc86386275)

[3. Expectation-maximization (EM) algorithm workflow and specific steps 5](#_Toc86386276)

[3.1 Plot sample’s KDE or PDP 5](#_Toc86386277)

[3.2 Run EM algorithm 6](#_Toc86386279)

[4. Additional Functions 8](#_Toc86386280)

[4.1 Plot confidence ellipse 8](#_Toc86386281)

[4.2 Calculate KL divergence 9](#_Toc86386282)

# 1. Introduction

***DzageEM*** is a MATLAB-based code that can be accessed via an executable file graphical user interface (GUI). It is designed to quantify the major age peaks, age ranges, and weights for a set of detrital geochronology data based on the Expectation-Maximization (EM) algorithm. The normal mixture model estimated by the EM algorithm can not only serve as a new detrital data visualization tool but can be used to be the alternative basis of KDE to calculate existing metrics. In addition, this code also can calculate the new quantitative metric of sample comparison in detrital geochronology, Kullback-Leibler divergence. The workflow is organized sequentially so that users can easily generate results (Fig. 1).

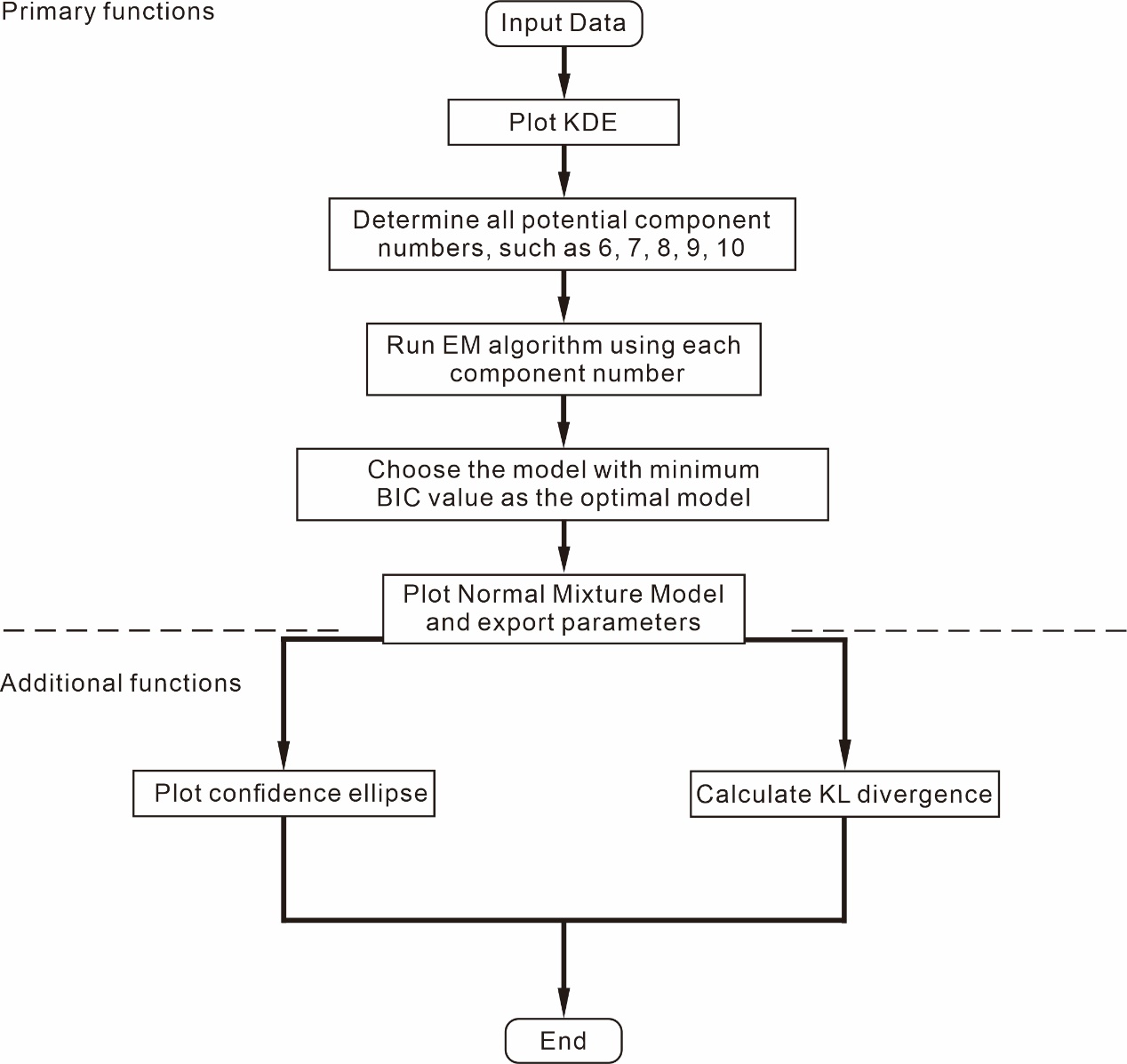


Figure 1. The workflow for estimating the major age parameters and calculating the KL divergence

# 2. Data import format

Data should be organized into a .xlsx file without header rows. When estimating a sample’s major age peaks and age ranges, each file is represented by a pair of columns where the first column is the mean ages, and the second column is the uncertainty associated with each age (1σ sigma). Please refer to Figure 2. When calculating the KL divergence between two samples, each file consists of four columns. The first and second columns are the ages and the uncertainty associated with each age (1σ sigma) of the first sample. The third and fourth columns are the ages and the uncertainty associated with each age (1σ sigma) of the second sample. The specific format is shown as Figure 3.

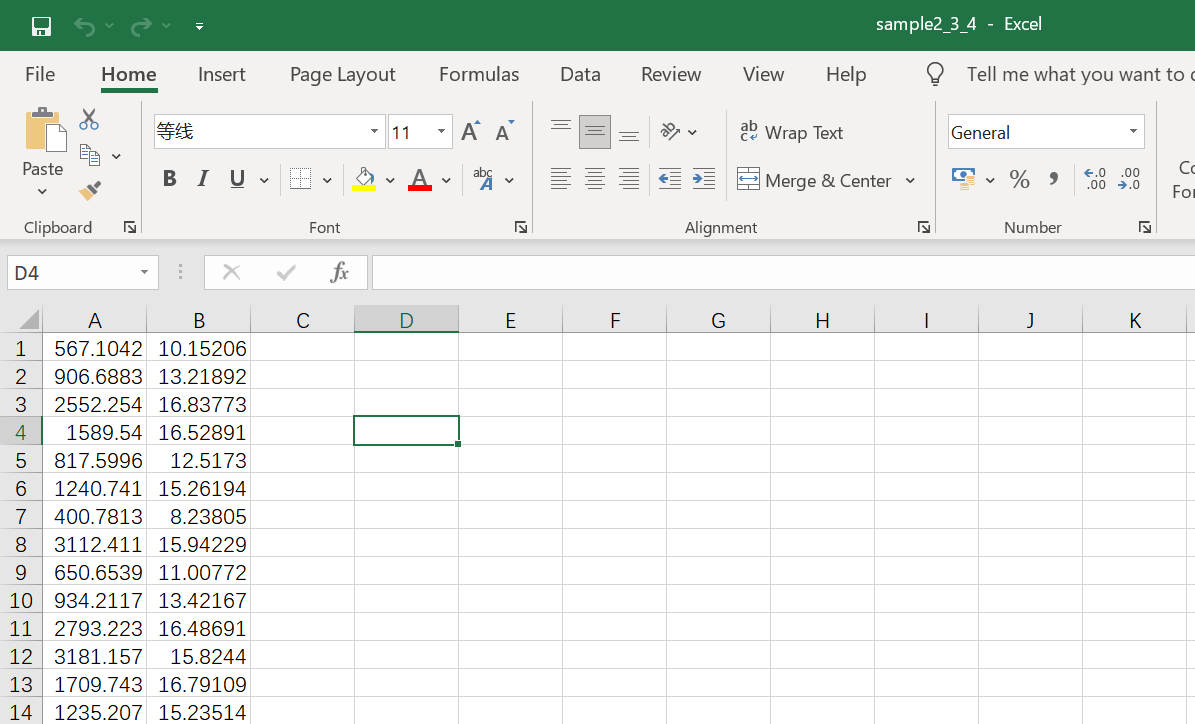


Figure 2. The data format for estimating a sample’s age parameters

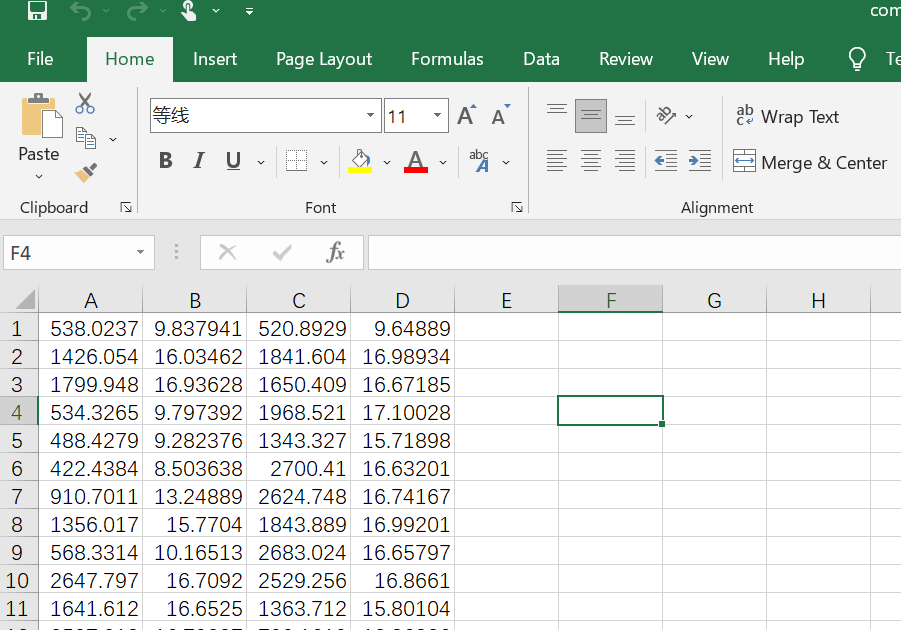


Figure 3. The data format for calculating two samples’ KL divergence

# 3. Expectation-maximization (EM) algorithm workflow and specific steps

## 3.1 Plot sample’s KDE

The EM algorithm needs a set of initial ages for further iteration, and more importantly, the initial ages cannot be selected arbitrarily. This paper proposed a new method for choosing initial values to guarantee the correct running of the EM algorithm. Researchers only need to input the expected component number, the application will select proper initial values automatically, and researchers can get the corresponding normal mixture model. Researchers can get all potential component numbers based on KDE with small bandwidth, keep trying using different component numbers and get multiple models. In general, the model whose BIC is minimum is preferred to be chosen. Therefore, the first step is to plot KDE with small bandwidths to select potential component numbers.

(1) Open the GUI. There are two ways to run the program. First, double click ‘**DzageEM.exe**’; Second, load the code named ‘**DzageEM.m**’ in MATLAB.

(2) In the pop-up window, click ‘**Input data**’ and import the prepared .xlsx file.

(3) Enter a fixed bandwidth for KDE, a small bandwidth (such as 20, 30, and 35) is suggested.

(4) Click ‘**Plot PDP and KDE**’ to generate the KDE and PDP.

(5) Observe the PDP and KDE, mainly KDE, guess possible component numbers. For example, Figure 4 indicates that this sample may have 9, 10, 11, 12, 13, 14, or 15 components.

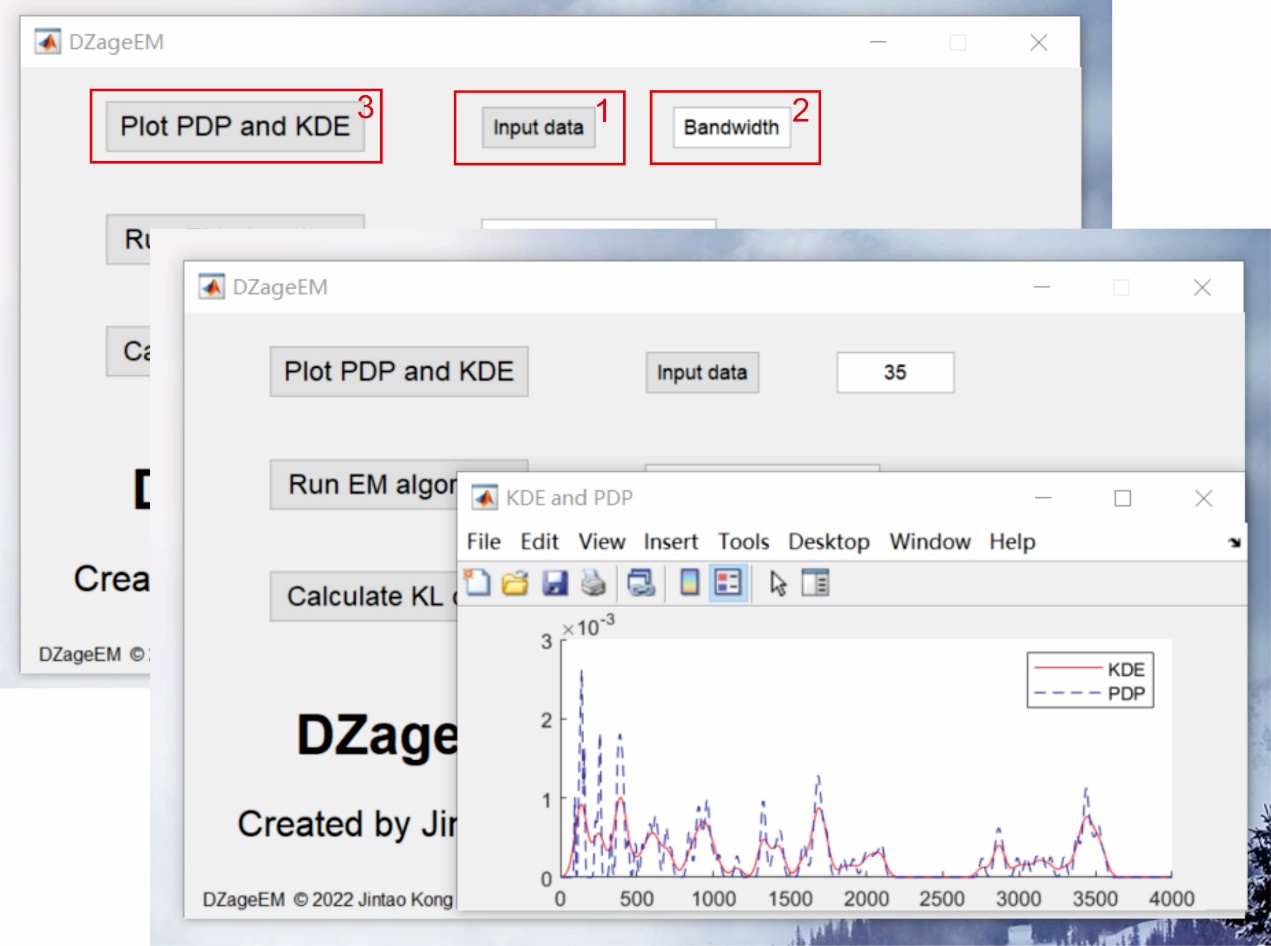


Figure 4. The specific steps of plotting sample’s KDE and PDP

## 3.2 Run EM algorithm

When obtaining all potential numbers of components, we can run the EM algorithm under the condition of each component number, to estimate different normal mixture models. For each model, their BIC will be calculated. In general, we prefer to choose the model whose BIC is minimum. The optimal model can provide age parameters and weights for researchers. The specific steps are as follows (Fig. 5):

(1) Build on the previous step, please input the expected component number, and click ‘**Run EM algorithm**’.

(2) In the pop-up window, users can get age peaks, age ranges, and corresponding weights.

(3) In the pop-up window, click ‘**Plot normal mixture model**’, users can get the sample’s normal mixture distribution.

(4) Go back to the first step, and run the EM algorithm using another component number (Fig. 6). Keep repeating until all component numbers are tried out. In general, the model whose BIC is minimum is preferred to be chosen, whereas, prior knowledge and analytical results should be considered comprehensively.

(5) When the bandwidth is too small, such as 10 or 15, there will be excessive age peaks, which may result in the EM algorithm running too long. After clicking ‘**Run EM algorithm**’, a time bar showing the degree of program running will pop up (Fig. 7). When the running time is too long, users can click ‘**Cancel**’ to stop the program from running. Users can increase bandwidth (such as 25, 30, or 35) to speed up running times.

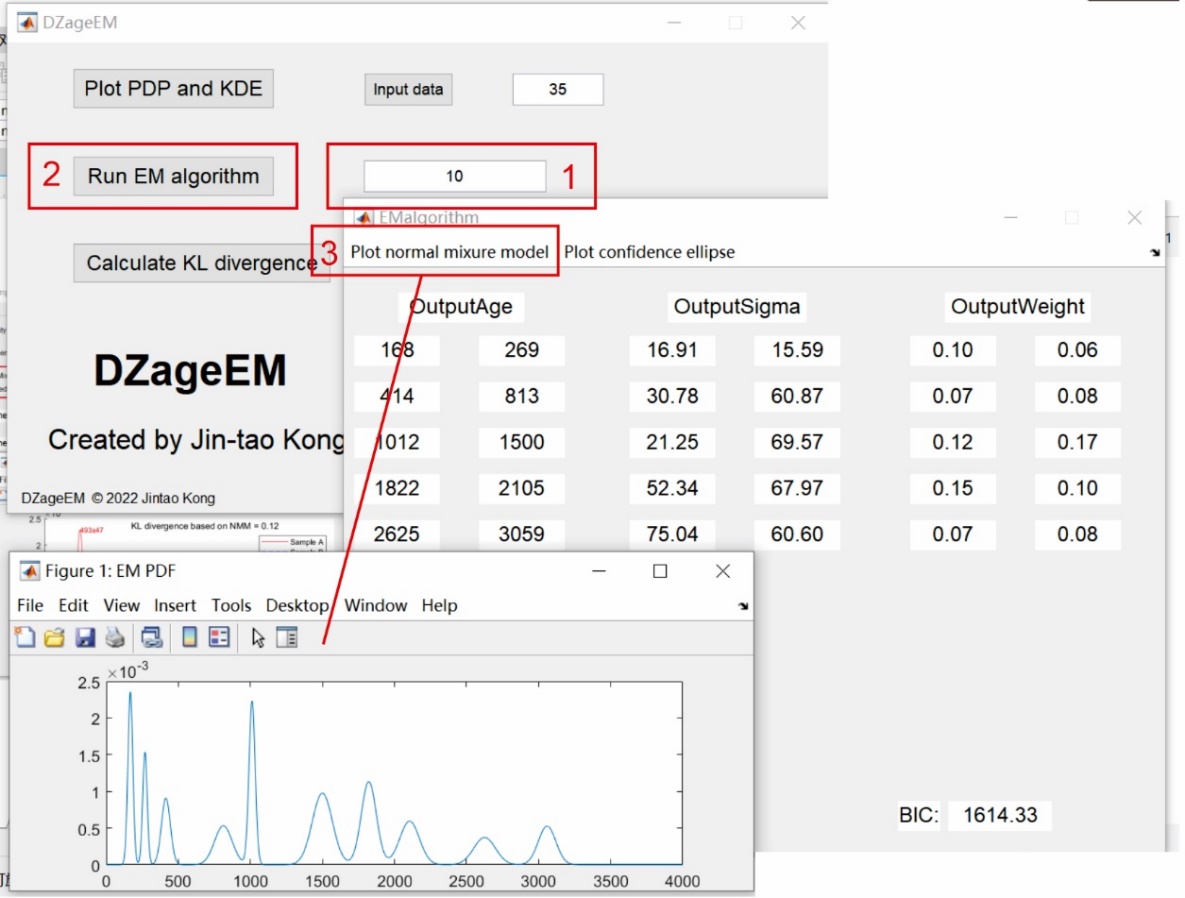


Figure 5. The specific steps of running EM algorithm when expected component number is 10.

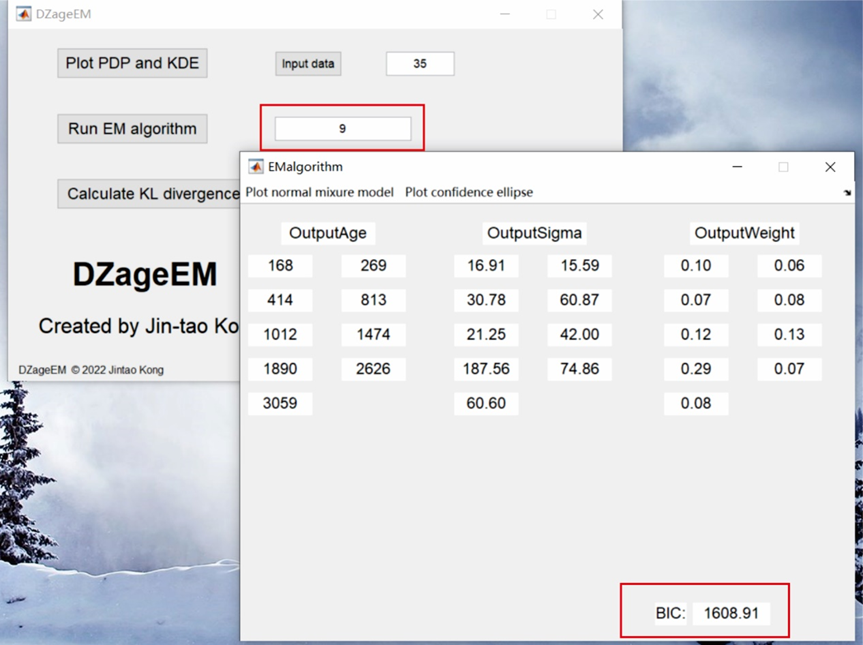


Figure 6. The specific steps of running EM algorithm when expected component number is 9.

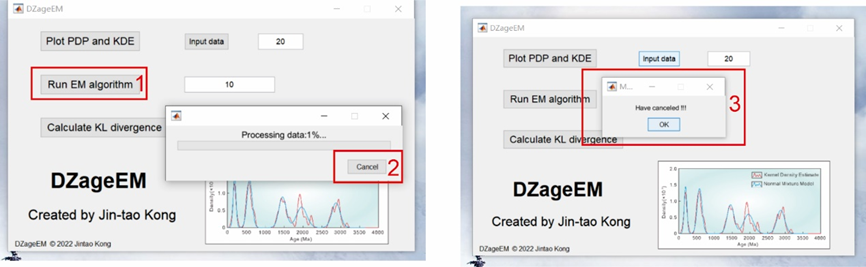


Figure 7. The specific steps of canceling the EM algorithm

# 4. Additional Functions

## 4.1 Plot confidence ellipse

Due to the effect of sampling error, the estimated parameters cannot equal to true values. Researchers can draw the confidence ellipse to determine the confidence interval of age peaks and weights to enhance the reliability of weights and age peaks.

(1) After running the EM algorithm, click ‘**Plot confidence ellipse**’ in the EM algorithm interface (Fig. 8) to draw the 95%-confidence ellipses of age peaks and weights (Fig. 8).

(2) Click ‘**Export**’ in the child window of ‘Confidence ellipse’, and the image can be exported as a .png, .bmp, .jpeg, or .eps file.

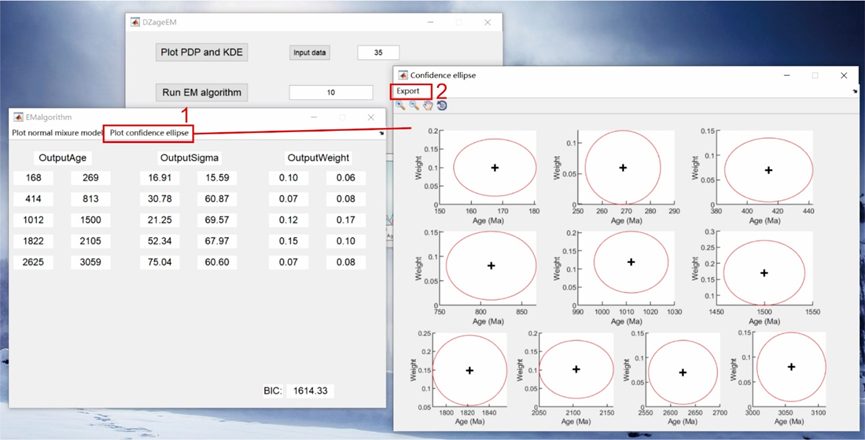


Figure 8. The specific steps of plotting confidence ellipse

## 4.2 Calculate KL divergence

This paper also introduces the KL divergence as a new quantitative metric. In the main window, users can click ‘Calculate KL divergence’, and a child window named ‘**KLdivergence\_Compare**’ will pop up. This application provides three options to calculate KL divergence values, which are based on the PDP, KDE, and normal mixture model (Figure 9).

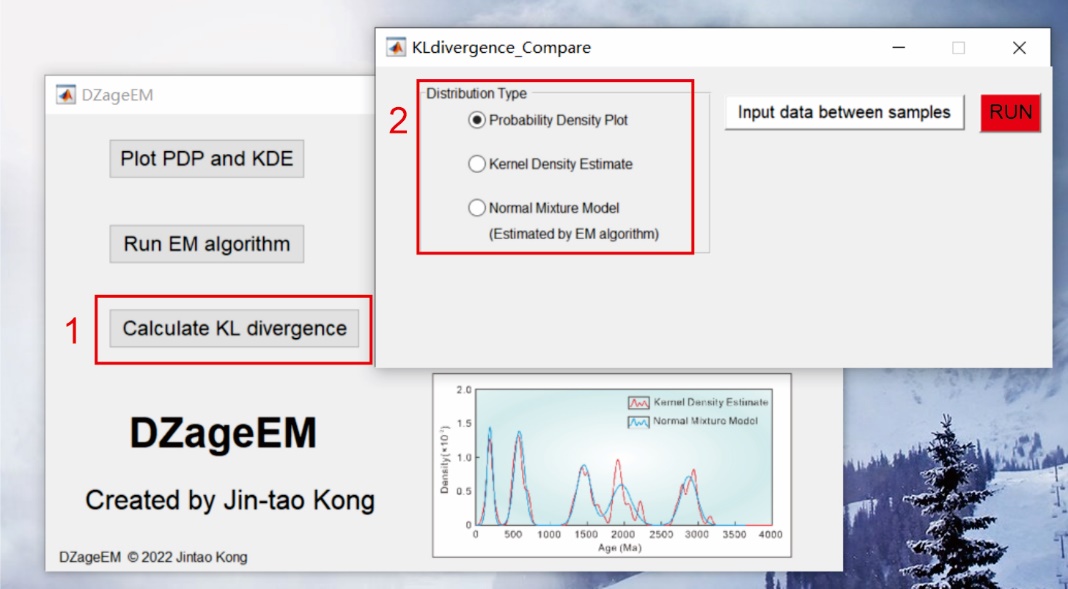


Figure 9 The child window for calculating KL divergence

(1) When users want to calculate the KL divergence value based on the PDP between two samples, they would click the radio button named ‘**Probability Density Plot**’. Then, users can input data files by clicking ‘**Input data between samples**’. The specific format of the data file is shown in Figure 3. Finally, PDPs and KL divergence values based on the PDP of two samples can be obtained by clicking ‘**RUN**’ (Figure 10). **HOWEVER, this paper strongly suggests that researchers should abandon this approach due to its poor theoretical foundations and performance.**

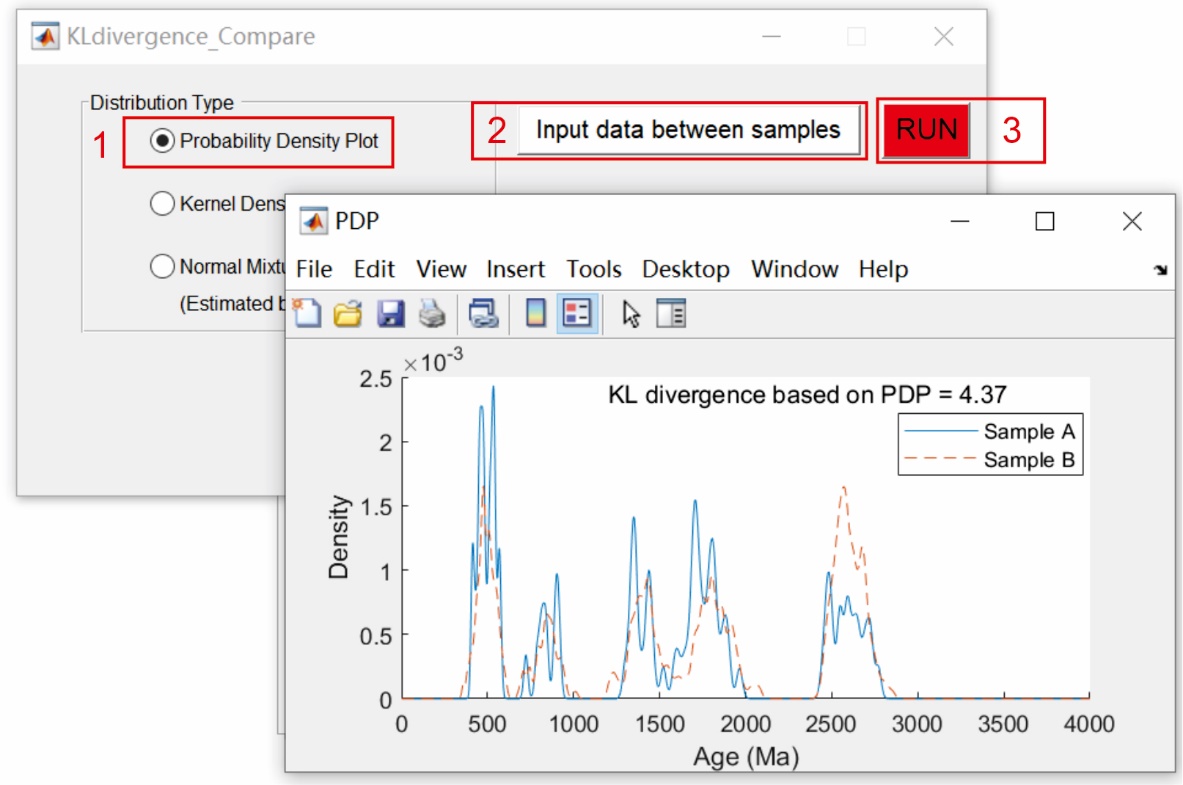


Figure 10 The specific steps of calculating KL divergence based on PDP

(2) When users want to calculate the KL divergence value based on the KDE between two samples, they would click the radio button named ‘**Kernel Density Estimate**’. Then, users can input data files by clicking ‘**Input data between samples**’. The specific format of the data file is shown in Figure 3. Additionally, users also need to input the bandwidths of Sample A and Sample B. Finally, KDEs and KL divergence value based on the KDE of two samples can be obtained by clicking ‘**RUN**’ (Figure 11).

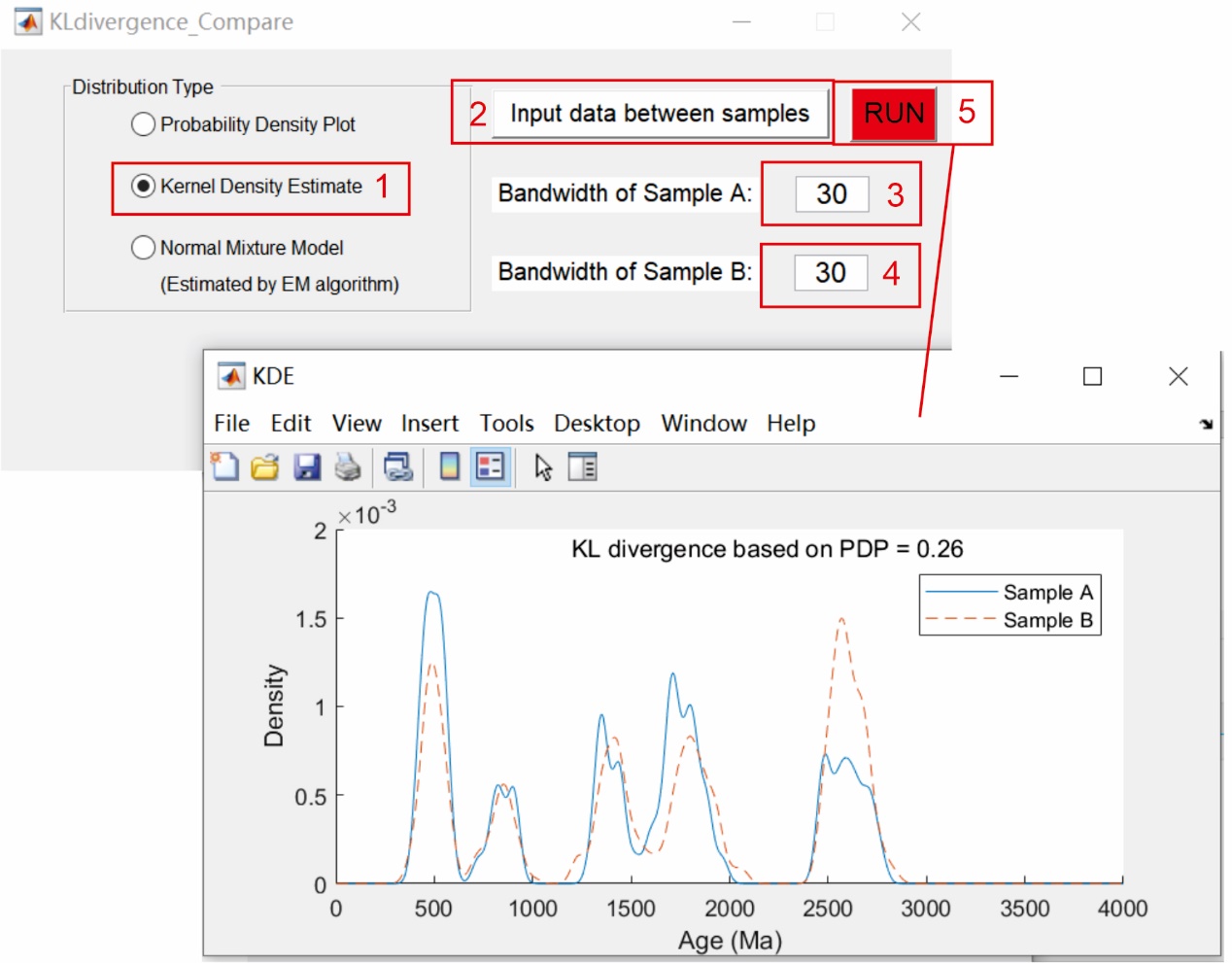


Figure 11. The specific steps of calculating KL divergence based on KDE

(3) When users want to calculate the KL divergence value based on the normal mixture model between two samples, they would click radio button named ‘**Normal Mixture Model**’. As with the above steps, users would input data files by clicking ‘**Input data between samples**’ and input the bandwidths of Sample A and Sample B. Next, users would input the expected component numbers of Sample A and Sample B, which can be obtained using the method introduced in Section 3.2. Finally, the normal mixture models and KL divergence value based on the NMM of two samples can be obtained by clicking ‘RUN’ (Figure 12).

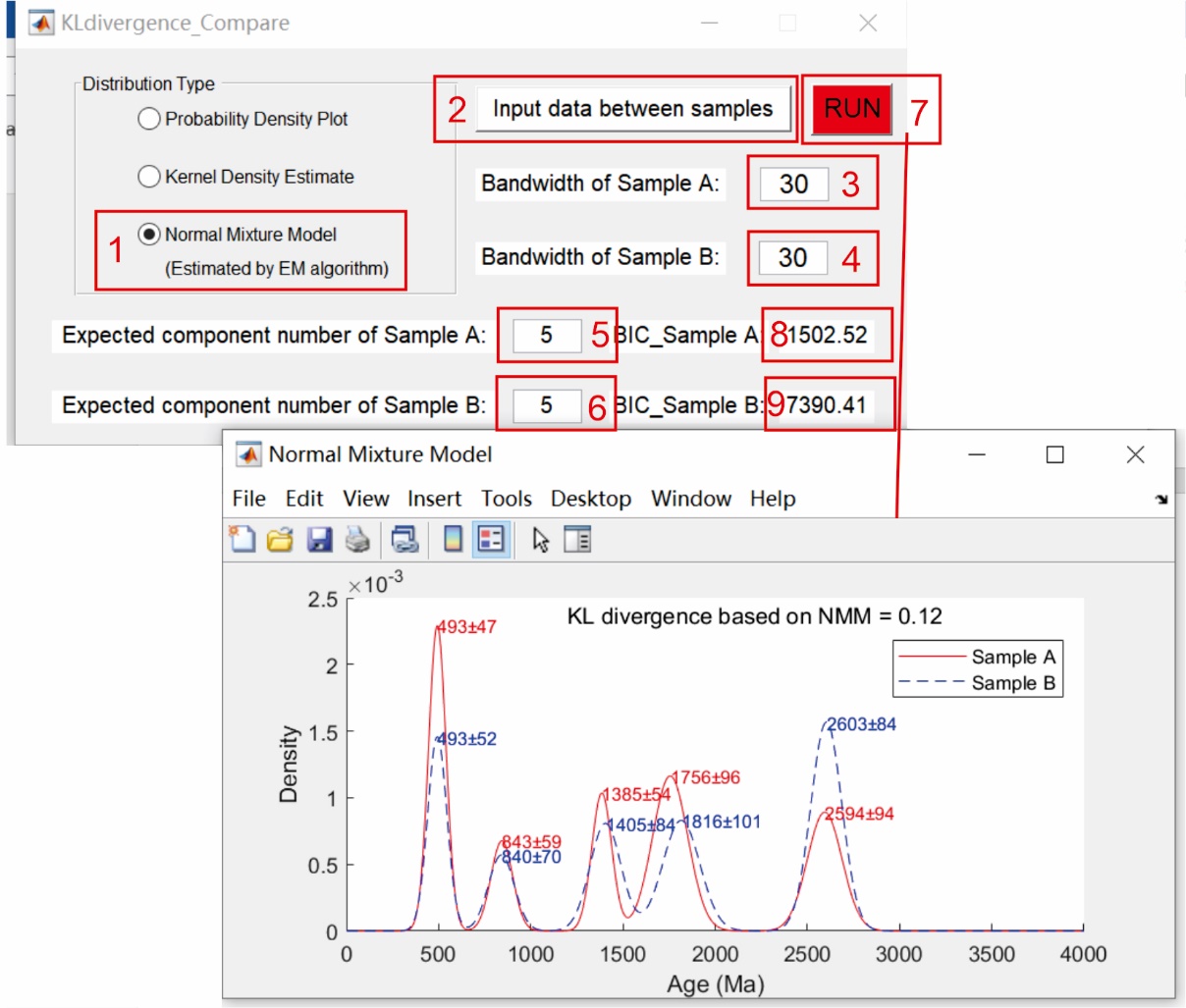


Figure 12. The specific steps of calculating KL divergence based on the normal mixture model