Guide to inferring Probability Model Ensembles (PMEs) for detrital zircon data and calculating Bayesian Population Correlation (BPC)

Workflow Overview:

1) makePME_GUI.m – use to infer PMEs for a set of detrital zircon age samples.
2) BPCunc_GUI.m – use to estimate the uncertainties on BPC values calculated between sample pairs.
3) evalBPC_GUI.m – use to plot the results of the BPC calculation, including estimated uncertainties.

Additional) PMEplot_GUI.m – use to plot PMEs.

Additional) BPC2frac_GUI.m – use to infer the overlapping proportions of two populations from their BPC value.

Dependencies

makePME_GUI.m and makePME.m require the MATLAB Global Optimization toolbox, as well as a script, nearestSPD.m, which is copyright (c) 2013, John D'Errico and provided according to the license text contained in "nearestSPD_license.txt", distributed with the collection of scripts. We also use a script, parfor_progressbar.m, which is copyright (c) 2016, Daniel Terry, and provided according to the license text contained in "parfor_progressbar_license.txt".

Tested hardware

This software has been successfully used on:

2014 Macbook Pro, 2.2 GHz, 4-core processor and 16 GB RAM, MATLAB_R2017b, macOS 10.12.6

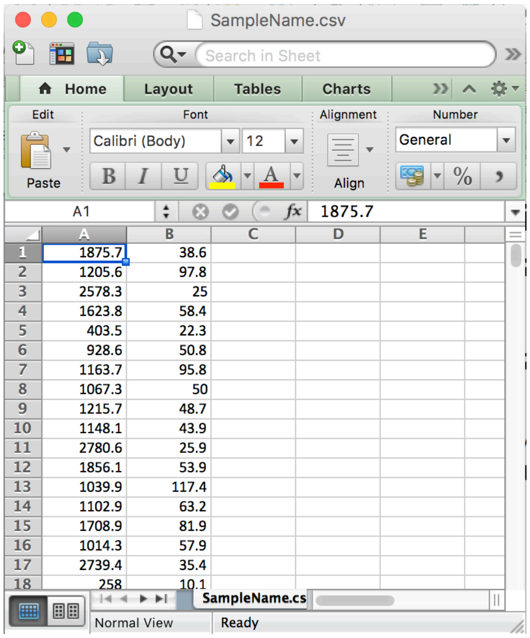
HP Windows 10 machine, 12-core processor and 64 GB RAM, MATLAB_R2016b.

Workflow Detail:

Note: For this demonstration, we run our collection of scripts on the 4 random subsamples (from data of Pullen et al., 2014, and Thomson et al., 2017), included with the scripts in the folder 'sample_data/'. If the scripts are run on these data, the results should resemble those found in this document.

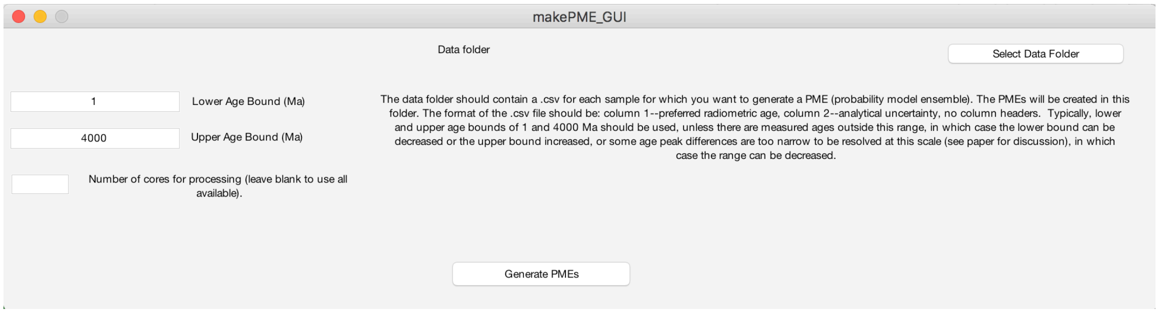
1. Inferring Probability Model Ensembles (PMEs) using makePME_GUI.m

For the samples you want to model, save the best ages of these samples, along with analytical uncertainties, in two-column .csv files, as follows:



	A	B	C	D	E
1	1875.7	38.6			
2	1205.6	97.8			
3	2578.3	25			
4	1623.8	58.4			
5	403.5	22.3			
6	928.6	50.8			
7	1163.7	95.8			
8	1067.3	50			
9	1215.7	48.7			
10	1148.1	43.9			
11	2780.6	25.9			
12	1856.1	53.9			
13	1039.9	117.4			
14	1102.9	63.2			
15	1708.9	81.9			
16	1014.3	57.9			
17	2739.4	35.4			
18	258	10.1			

Ensure that the .csv files corresponding to all desired samples are located in a single folder, with no other .csv files. I suggest creating a new folder. Open and run 'makePME_GUI.m' in MATLAB. You should see the following window appear:



1 Lower Age Bound (Ma)

4000 Upper Age Bound (Ma)

Number of cores for processing (leave blank to use all available).

The data folder should contain a .csv for each sample for which you want to generate a PME (probability model ensemble). The PMEs will be created in this folder. The format of the .csv file should be: column 1--preferred radiometric age, column 2--analytical uncertainty, no column headers. Typically, lower and upper age bounds of 1 and 4000 Ma should be used, unless there are measured ages outside this range, in which case the lower bound can be decreased or the upper bound increased, or some age peak differences are too narrow to be resolved at this scale (see paper for discussion), in which case the range can be decreased.

Click the "*Select Data Folder*" button to select the folder where your .csv files are located. The path of this folder will then appear next to the button. Change the age bounds if desired, and click "*Generate PMEs*". You may also optionally specify how many cores to use to run this script. After you click *"Generate PMEs"*, you may not see any activity in the MATLAB Command Window, although the parallel pool should start. Almost immediately, you should see a progress bar that says, *"Inferring PMEs…"*. Generating PMEs typically takes ~5-10 minutes per sample and depends in part on sample size, with one very large sample having taken several hours. These times are for a 2014 Macbook Pro with 2.2 GHz, 4-core processor and 16 GB RAM.

File architecture (optional to read): As the makePME script runs, it generates two .csv files per sample in the selected folder, plus two additional .csv files for each possible pairing of samples in the selected folder. For a given sample or sample pair, one of the two files contains a PME (a Markov chain), and is named with the sample name followed by 'chain.csv'. Each row of the 'chain.csv' file corresponds to a different probability model that has been accepted into the PME. The columns store the values of the 50 model parameters of each of these probability models. The second of the two files per sample or sample pair contains a list of the log likelihood values of the PME inferred for that sample or sample pair. This second file is named with the sample name followed by 'logLk.csv'. The 'logLk.csv' file is a single column, where each row contains a log likelihood value that corresponds with the model in the same row of the 'chain.csv' file. These files appear in a subdirectory of the folder containing the sample .csv files named 'chains/'.

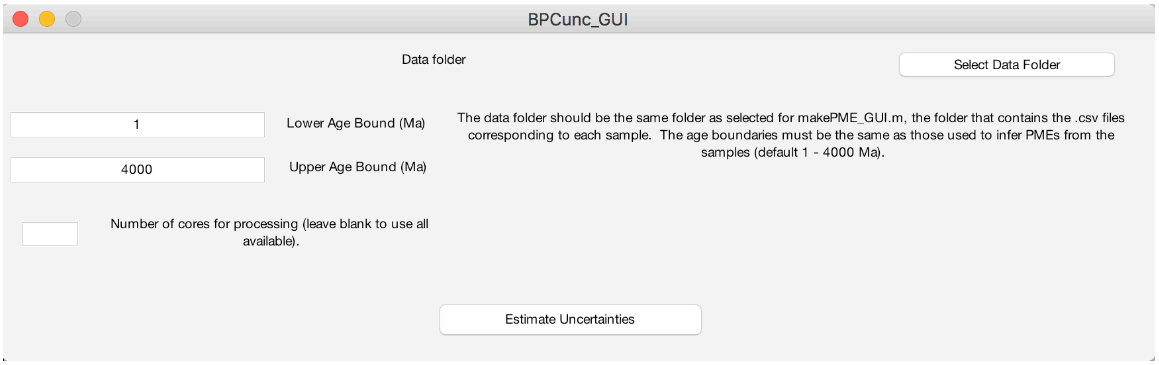
In addition, log files are generated during each run and are stored in a 'log/' folder. The filenames of the log files match the filenames of the corresponding sample .csv files with "log.txt" appended.

The files resulting from the comparison of two samples are named as above except the names of the two samples are concatenated such that the filenames read "SampleName1_SampleName2chain.csv", where "SampleName1" and "SampleName2" are the filenames of the sample .csv files and "chain.csv" could also be "logLk.csv" or "log.txt".

Note on storage of PME information: At this time, the domain over which modeling is conducted is not saved with the PME models and likelihoods, so the user must be sure to be consistent in defining their domain. We recommend a domain of 1 to 4000 Ma, unless otherwise is necessary. All modeling is done in log age space, as discussed in the paper text.

2. Estimating BPC uncertainties with BPCunc_GUI.m

Following the inference of PMEs using ‘makePME_GUI.m’, an additional phase of simulation is undertaken to estimate the uncertainties on resulting BPC values (see paper text). Open and run ‘BPCunc_GUI.m’ in MATLAB, and you should see the following window appear:



BPCunc_GUI

Data folder

Lower Age Bound (Ma): 1

Upper Age Bound (Ma): 4000

The data folder should be the same folder as selected for makePME_GUI.m, the folder that contains the .csv files corresponding to each sample. The age boundaries must be the same as those used to infer PMEs from the samples (default 1 - 4000 Ma).

Number of cores for processing (leave blank to use all available).

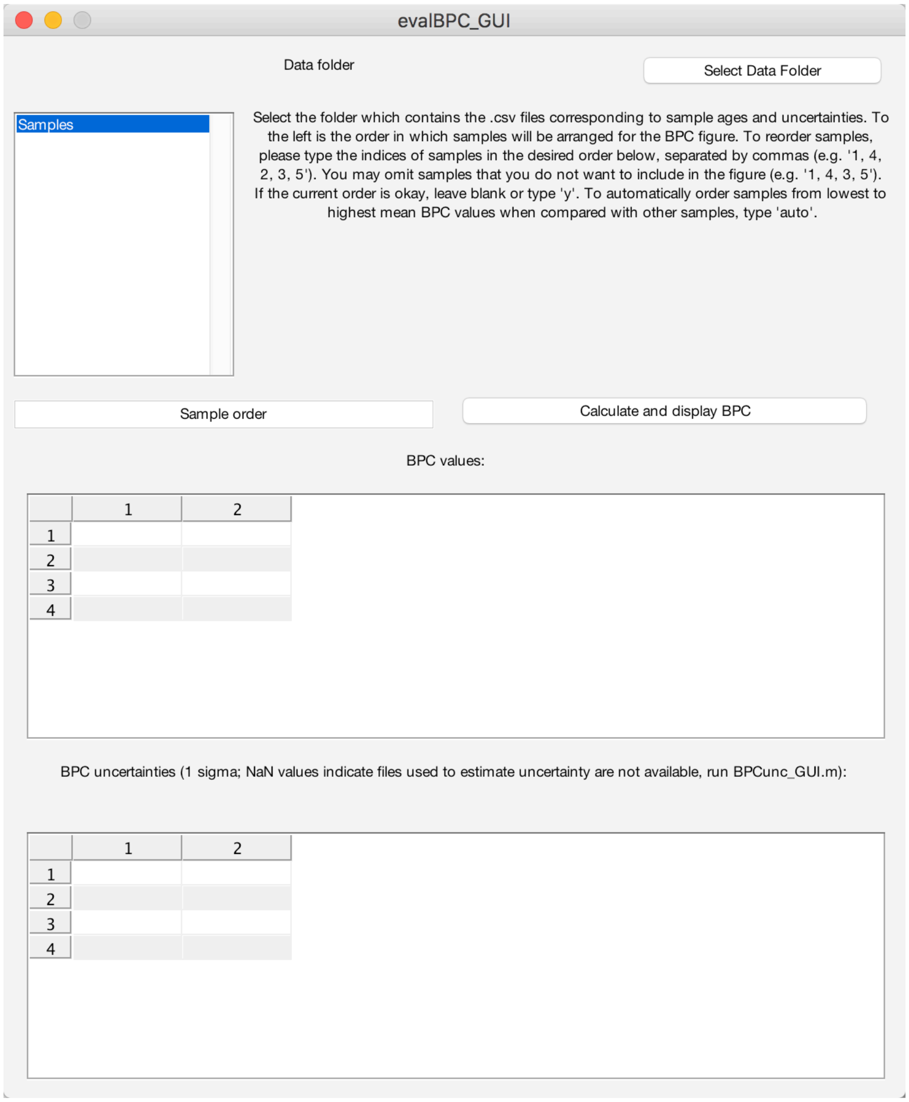
Press *“Select Data Folder”*, and select the same folder you selected in ‘makePME_GUI.m’, which contains the .csv files corresponding to sample ages. Use the same age bounds as in ‘makePME_GUI.m’, and optionally change number of cores to use to run the script. Click *“Estimate Uncertainties”*. You should see a progress bar appear that says *“Estimating uncertainties...”*. This procedure takes somewhat less time than ‘makePME_GUI.m’.

File architecture (optional to read): ‘BPCunc_GUI.m’ creates an additional subdirectory in the folder you selected, called ‘unc/’. In this folder, one additional .csv file is created for each sample along with one .csv file for each pair of samples. The files created for individual samples contain simulated age samples obtained by resampling the PME for each sample, along with the likelihood value of the maximum likelihood model for that simulated sample (see text for discussion). These files end with ‘resample.csv’ and each column corresponds to a single simulated sample. Column headers are the maximum likelihood values of each simulated sample. The files created for pairs of samples contain the likelihood values of the maximum likelihood model for pairs of simulated samples, and end with ‘jointML.csv’. These files are a single column of maximum likelihood values. Like ‘makePME_GUI.m’, these files do not record the domain used for modeling, so care must be paid.

3. Calculating Bayesian Population Correlation (BPC) values using evalBPC_GUI.m

Calculating BPC values requires that PMEs have been generated and that the directories created during the execution of the makePME script have not been moved.

Run the ‘evalBPC_GUI.m’ script. You should see the following window:



Select the folder which contains the .csv files corresponding to sample ages and uncertainties. To the left is the order in which samples will be arranged for the BPC figure. To reorder samples, please type the indices of samples in the desired order below, separated by commas (e.g. '1, 4, 2, 3, 5'). You may omit samples that you do not want to include in the figure (e.g. '1, 4, 3, 5'). If the current order is okay, leave blank or type 'y'. To automatically order samples from lowest to highest mean BPC values when compared with other samples, type 'auto'.

BPC values:

	1	2
1		
2		
3		
4		

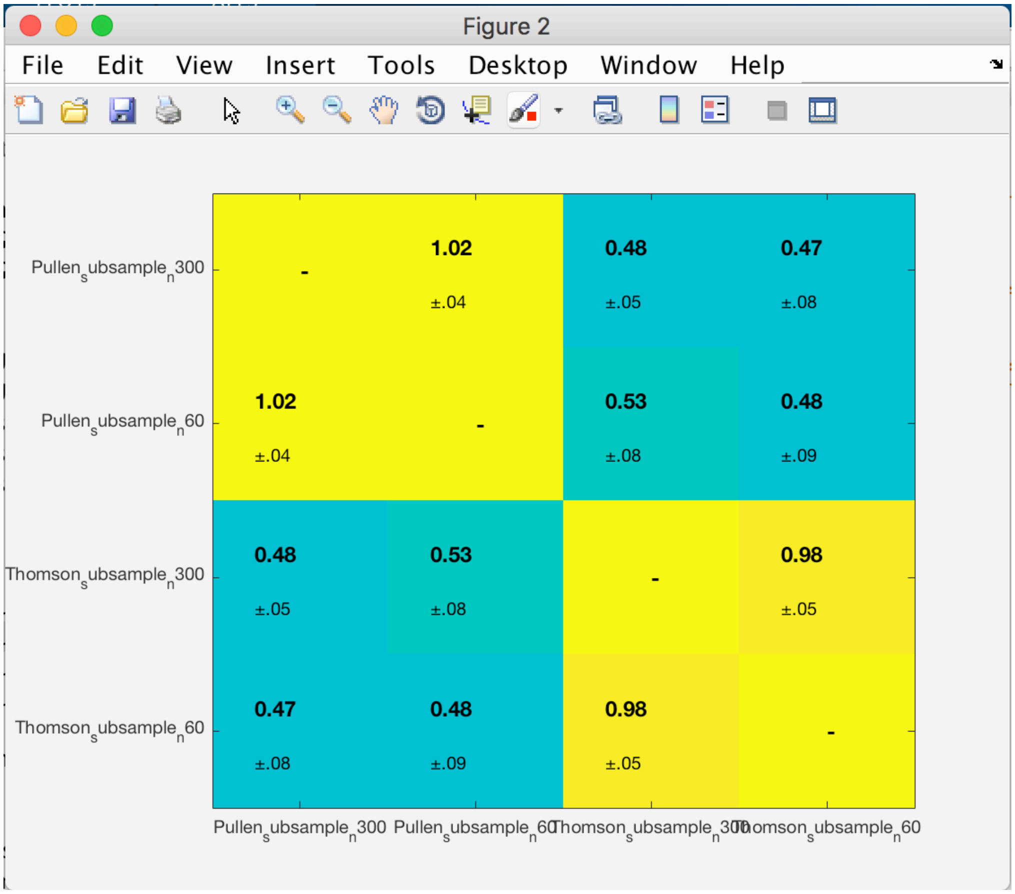
BPC uncertainties (1 sigma; NaN values indicate files used to estimate uncertainty are not available, run BPCunc_GUI.m):

	1	2
1		
2		
3		
4		

As before, click the “*Select Data Folder*” button and choose the folder where your sample .csv files are located. The path of the selected folder will appear next to the button. In addition, the filenames of the sample .csv files will be shown in the listbox on the left side of the window.

The BPC values, once calculated, will be shown in an NxN matrix, where N is the number of compared samples. The "*Sample order*" textbox allows you to change the order in which the samples will appear in this matrix. To specify the order, type the indices of the sample filenames (shown in the listbox) in the desired order, separated by commas (e.g. '1, 3, 4, 2, 5'). See the text in the window for further instructions. You can also leave "*Sample order*" blank or enter "y" to use the default order shown in the listbox, or you can type "auto" to automatically order the samples in terms of lowest to highest mean BPC value calculated with all other samples.

Click "*Calculate and display BPC*". You should see a color-coded table output to a new figure:

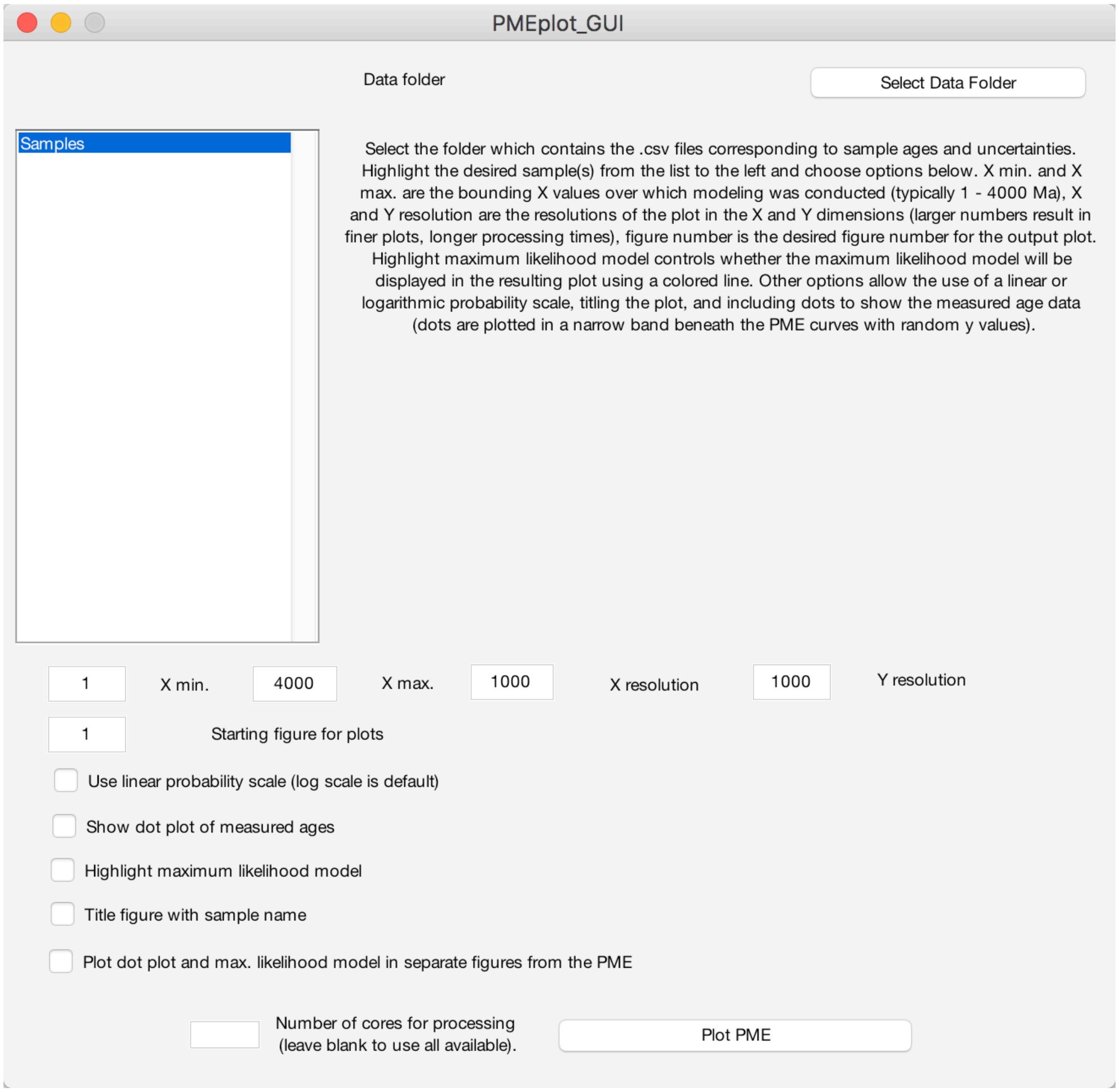


	Pullen$_s$ubsample$_n$300	Pullen$_s$ubsample$_n$60	Thomson$_s$ubsample$_n$300	Thomson$_s$ubsample$_n$60
Pullen$_s$ubsample$_n$300	-	1.02 ±.04	0.48 ±.05	0.47 ±.08
Pullen$_s$ubsample$_n$60	1.02 ±.04	-	0.53 ±.08	0.48 ±.09
Thomson$_s$ubsample$_n$300	0.48 ±.05	0.53 ±.08	-	0.98 ±.05
Thomson$_s$ubsample$_n$60	0.47 ±.08	0.48 ±.09	0.98 ±.05	-

This figure shows the BPC value and uncertainty for each pair of compared samples. The colors illustrate the BPC value on the MATLAB parula colormap stretched from 0 to 1. In addition, the "*BPC value*" and "*BPC uncertainties (1 sigma)*" fields in the window are populated with values that can be copied to the clipboard. If 'BPCunc_GUI.m' hasn't yet been used to estimate the BPC uncertainties, uncertainties will not be shown.

Additional. Displaying Probability Model Ensembles (PMEs) using PMEplot_GUI.m

Run the ‘PMEplot_GUI.m’ script. You should see the following window:



Select the folder which contains the .csv files corresponding to sample ages and uncertainties. Highlight the desired sample(s) from the list to the left and choose options below. X min. and X max. are the bounding X values over which modeling was conducted (typically 1 - 4000 Ma), X and Y resolution are the resolutions of the plot in the X and Y dimensions (larger numbers result in finer plots, longer processing times), figure number is the desired figure number for the output plot. Highlight maximum likelihood model controls whether the maximum likelihood model will be displayed in the resulting plot using a colored line. Other options allow the use of a linear or logarithmic probability scale, titling the plot, and including dots to show the measured age data (dots are plotted in a narrow band beneath the PME curves with random y values).

1 X min. 4000 X max. 1000 X resolution 1000 Y resolution

1 Starting figure for plots

☐ Use linear probability scale (log scale is default)

☐ Show dot plot of measured ages

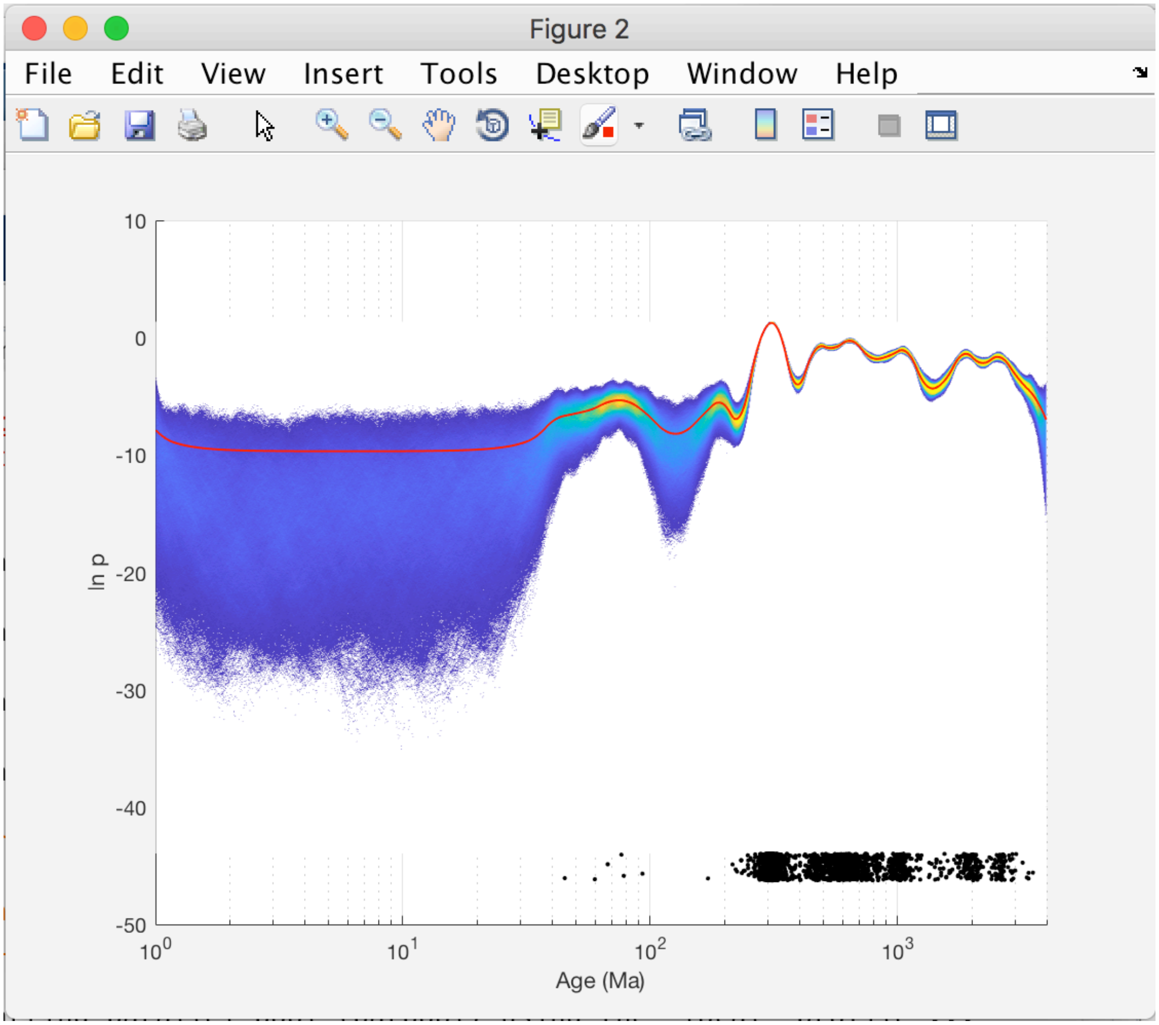
☐ Highlight maximum likelihood model

☐ Title figure with sample name

☐ Plot dot plot and max. likelihood model in separate figures from the PME

Number of cores for processing
(leave blank to use all available).

Again use “*Select Data Folder*” to select the folder where the sample .csv files are stored. As with evalBPC_GUI.m, the folder path will be shown next to the button and the samples contained in the folder will be shown in a listbox on the left. Highlight the desired sample(s) in the listbox and specify options using the text boxes and check boxes. Options that can be specified in textboxes include the x domain over which modeling was conducted (this is *not* just the desired bounds of the plot; use the same x min and x max values as for ‘makePME_GUI.m’ and ‘BPCunc_GUI.m’), and x and y resolution for the plot, and the number of the figure for output. In addition, checkboxes allow you to specify whether to highlight the maximum likelihood probability model in the PME plot, whether to title the plot with the sample name, whether to use a linear or logarithmic probability scale, and whether to include a dotplot of measured ages in the figure. The dotplot appears in a thin band beneath the probability models, similar to plots shown in Pullen et al. (2014). The dotplot and maximum likelihood model can be plotted in their own figure windows, as well, which can be useful if the further processing of the figures is intended in vector-based graphics software. The number of cores for the operation can also be specified. Once the desired options are specified, click "*Plot PME*". A progress bar should appear. Once reading and processing the data is complete, the resulting plot(s) will appear:

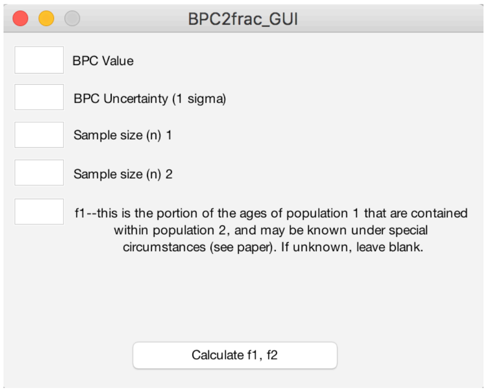


In this example, "Highlight maximum likelihood model" was selected, along with "Show dot plot of measured ages", and all other check boxes were left blank. The x and y resolutions were set to 1000. This figure shows a natural logarithmic age scale. In this script, the area covered by the probability models of the PME is discretized into cells in the x and y directions with the number of cells in one dimension equal to the resolution input into the GUI. Then, the cells that have probability models that pass through them are colored according to the number of models that pass through them, using the MATLAB parula colormap.

If more than one sample was selected in the listbox in the main window, the PME plots for each selected sample will appear sequentially.

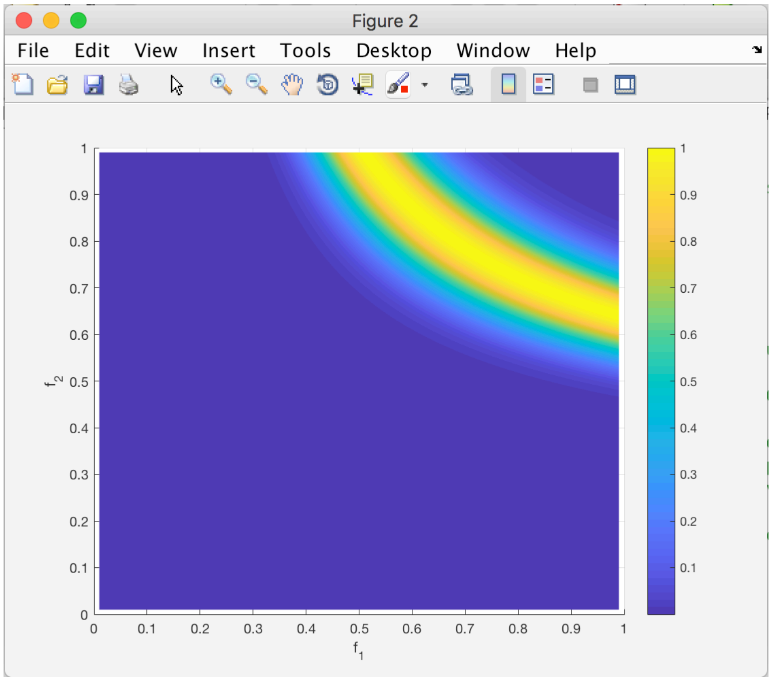
Additional. Inferring the shared fractions of two populations from BPC values using BPC2frac_GUI.m

BPC values have a functional relationship to the shared fraction of two detrital zircon populations (the fraction of age peaks of each population that is shared with the other population), which can be derived analytically (see Tye et al. text). Thus, a BPC value can non-uniquely constrain the shared fractions of both populations. This calculation is facilitated by the BPC2frac_GUI.m script. First, run BPC2frac_GUI.m. The following window should appear:



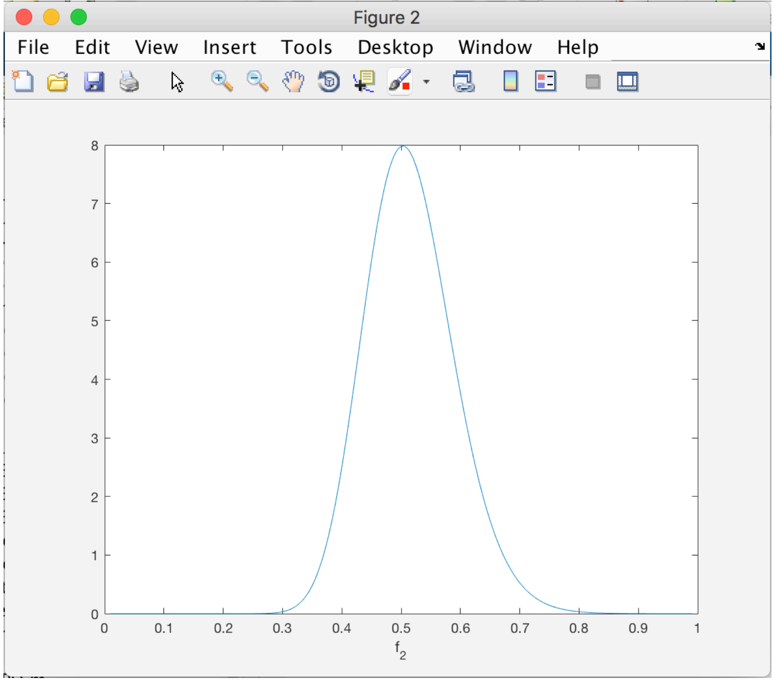
f1--this is the portion of the ages of population 1 that are contained within population 2, and may be known under special circumstances (see paper). If unknown, leave blank.

Fill out the text box fields, including BPC value and uncertainty, along with the sample sizes of the compared samples. In specific circumstances, the shared proportion of one sample may be assumed. For instance, if two samples are taken from two positions on the same river network such that the water and sediment that flow past the first sample location subsequently flow past the second sample location, then all the age peaks in the first sample can be assumed to be shared with the second sample, resulting in an f1 value of 1 (see text for discussion). If f1 can be assumed, then enter it as well. If no f1 value is entered, then the output will appear like this:



Here, colors indicate the relative likelihood of coordinate pairs of (f1, f2) values. These values are obtained by solving Eqn. C.7 in Appendix C in the text numerically for the given BPC value and uncertainty. This result was generated for a BPC value of 0.75, uncertainty of 0.05, and sample sizes of 100 and 300.

If f1 can be assumed, the output will appear like this:



Here, the plot shows the likelihoods of different values of f2 for the given value of f1. This example plot was generated using the same parameters as above, plus an f1 value of 1. In addition, the mean and standard deviation of this distribution are output to the MATLAB Command Window.

Notes

1. Use of scripts without GUIs

In this document, we describe the use of our scripts using GUI functions created for the main scripts, through which all functionality associated with PME inference and BPC calculation can be accessed. These GUIs merely input parameters into respective scripts that can be run from the Command Window or from within other scripts. These include makePME.m, BPCunc.m, evalBPC.m, PMEplot.m, and BPC2frac.m. Documentation for all of our scripts are included in in-line comments in the code.