

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

Table of Contents

1. Included files – a brief description of the contents of this package
2. Installation versus execution from MATLAB
3. Dependencies – the MATLAB versions and toolboxes necessary to run the scripts.
4. Workflow – step by step instructions for running the scripts
5. General Notes – useful information on running the scripts that doesn't fit in the step by step guide
6. Tested Hardware – specifications of machines on which the script has been run.

1. Included files

Included in this package are scripts necessary in order to infer PMEs, estimate BPC uncertainties, display calculated BPC values, display PME plots, and infer the shared proportions of two samples based on BPC values. In addition to being presented in their original form as MATLAB scripts, these files have been compiled into a standalone application. Installers for the application for Windows and Mac are also included. The main folder, which contains this document, contains a main menu .m file and accompanying .fig file:

'BPCmainmenu.m' – MATLAB script of the main menu for PME inference and BPC calculations, through which all included functionality can be accessed.

'BPCmainmenu.fig' – MATLAB file that contains information about the graphical user interface (GUI) for the main menu script.

The following subdirectories are also contained in this package:

'backend/' – contains scripts the user does not need to access directly. A short readme file in the folder discusses the function of each script. This folder also contains two third-party scripts, which are noted in the Dependencies section, below.

'sample_data/' – includes a set of sample data with which the scripts can be tested (see section Workflow, below).

The installers for the standalone applications (useful if the user lacks MATLAB or one of the required toolboxes—see Dependencies section below) are:

'BPCinstallerWin_web.exe' – this is the installer for Windows

'BPCinstallerMac_web.app' – this is the installer for Mac

Note that these installers require an Internet connection and were compiled using the MATLAB and operating system versions listed in the Tested Hardware section below.

The package also contains a text readme file, a text license file for the set of scripts, and this introductory document.

2. Installation versus execution from MATLAB

If the user lacks the required versions of MATLAB and the toolboxes used by the BPC scripts (see Dependencies section), then the BPC scripts can be installed as a standalone application. This requires running one of the two installers (*'BPCinstallerWin_web.exe'* for Windows machines, *'BPCinstallerMac_web.app'* for Mac) and clicking through the on-screen instructions. The installers were created using a built in function in MATLAB, and were compiled using the MATLAB and OS versions listed under the Tested Hardware section. An Internet connection is required for this installation because the installer downloads and installs the MATLAB runtime environment. Note that these standalone applications operate independently of the .m files contained in this package, and the directory into which you install the application shouldn't affect its functionality. The sample data that we use in the example Workflow section below will not be installed using the installer. In order to work with this data, the user must

make sure to download it in addition to the installer and then open the download location using the BPC application.

Note that if you run the standalone application instead of the MATLAB files, there will likely be times when you start the application or issue a command and the application appears not to do anything for several seconds. Despite these periods of seeming inactivity, the application is likely continuing as intended. In addition, please only click buttons once on the GUI and then wait for the function to execute.

On windows machines, the parallel computing performed by the scripts may cause the objection of Windows firewall. On Mac machines, note that the actual application file will be located at '*installed_directory*/application/BPC', where '*installed_directory*' is specified by the user during installation.

3. Dependencies

Note: Dependencies were assessed using the MATLAB function [matlab.codetools.requiredFilesAndProducts](#). The user can check which toolboxes they have installed in MATLAB by typing 'ver' into the command window.

The BPC scripts require the following MATLAB toolboxes (versions shown in parentheses):

- MATLAB (9.3)
- Optimization Toolbox (8.0)
- Statistics and Machine Learning Toolbox (11.2)
- Curve Fitting Toolbox (3.5.6)
- Parallel Computing Toolbox (6.11)
- Global Optimization Toolbox(3.4.3)

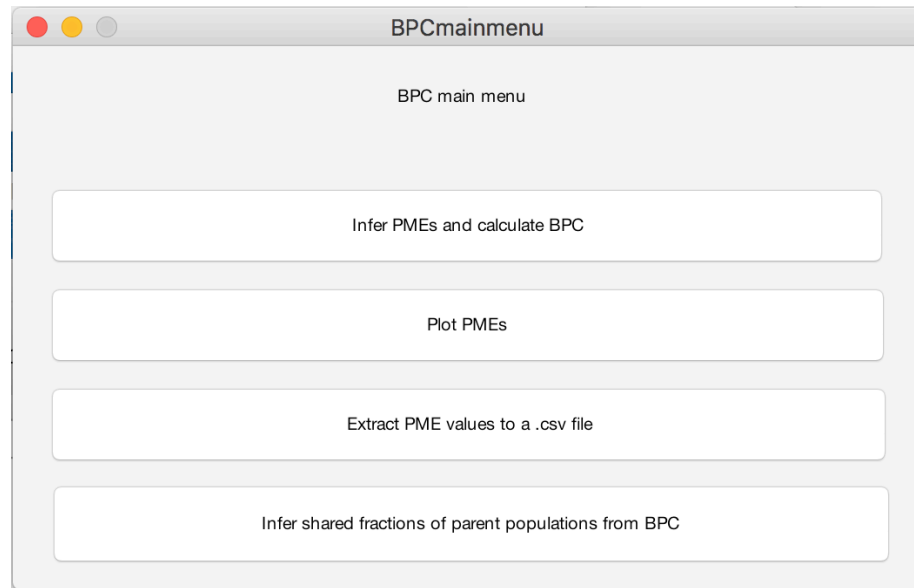
Two additional third-party scripts are required for our collection of scripts to function:

- nearestSPD.m*, which is copyright (c) 2013, John D'Errico, and is provided according to the license text contained in "nearestSPD_license.txt", distributed with our collection of scripts.
- parfor_progressbar.m*, which is copyright (c) 2016, Daniel Terry, and is provided according to the license text contained in "parfor_progressbar_license.txt".

4. Workflow

The first step of the workflow is to start the standalone application, or run the '*BPCmainmenu_GUI.m*' script from MATLAB (ensuring that your MATLAB

installation meets the version and toolbox requirements shown in the Dependencies section below). Upon doing so, you should see a simple window with buttons:



These four buttons allow access to the major functions included in this package by clicking the respective button. The remainder of this workflow includes 4 parts: A, B, C, and D, which correspond with the four buttons. Part A shows how to infer PME's and calculate BPC. Part B shows how to plot PME's. Part C shows how to extract probability values from the PME's to a .csv file. Part D shows how to infer the shared proportions of two zircon age populations from their BPC value.

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.

A. Infer Probability Model Ensembles (PME's), estimate BPC uncertainties, and display BPC values from a single GUI using BPConeclick_GUI.m.

For the samples you want to model, save the best ages of these samples, along with analytical uncertainties (1σ), 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			

The first column is the preferred measured age for each analysis in millions of years and the second column is the 1σ analytical uncertainty calculated for the preferred measured age.

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. Click the “*Infer PMEs and display BPC*” button from the main menu. You should see the following window appear:

BPConeclick_GUI

Data folder Select Data Folder

Age lower bound (Ma)

Age upper bound (Ma)

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

Sample order

Calculate and display BPC

BPC values:

	1	2
1		
2		
3		
4		

BPC uncertainties (1 sigma):

	1	2
1		
2		
3		
4		

Select the folder which contains the .csv files corresponding to sample ages and uncertainties. The correct format for these files is discussed in the introductory PDF. This "one-click" script will infer PMEs, estimate BPC uncertainties, and display BPC values for a set of samples. Input the upper and lower age bounds for modeling using the text boxes to the left. These should be set to 4000 and 1 respectively unless there is a compelling reason to do otherwise. Input the number of cores to use for processing.

Once the data folder is selected, this list box shows 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'.

Finally, click calculate and display BPC to begin the process. This can take a long time (see introduction document), but progress is saved as the script proceeds in case of interruption. Once the analysis is complete, the BPC values can be displayed again by running this script. The PME inference won't be repeated so the process will be quick.

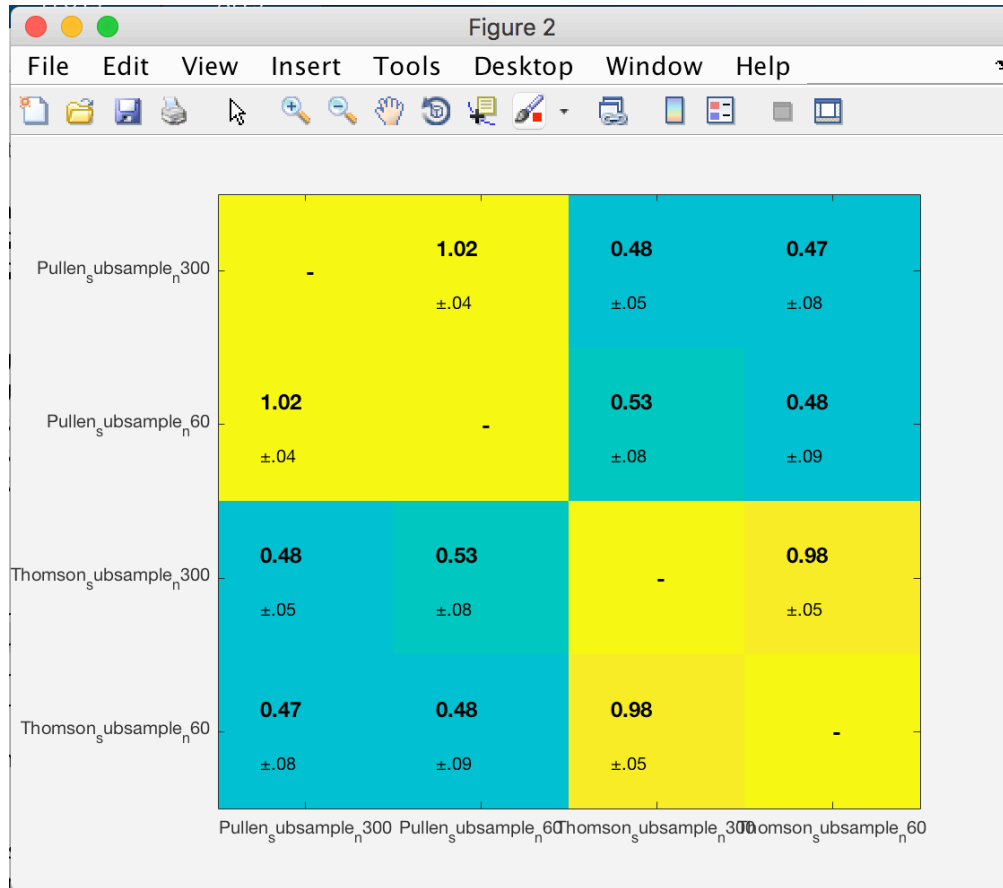
Click the “*Select Data Folder*” button to select the folder where your .csv files are located. The path of this folder will then appear next to the button. Change the age bounds if desired, though bounds of 1 Ma and 4000 Ma should generally be used unless there is a strong reason to use a different set of bounds. *BPC values calculated using a given set of bounds are not directly comparable to BPC values calculated using a different set of bounds.* Enter the number of cores you desire to be used for the BPC analysis. If this textbox is left blank, all available cores will be used.

If a folder has been selected, 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*” in order to begin the process of inferring PMEs, estimating BPC uncertainty, and ultimately displaying BPC values. First, the script will infer PMEs for the samples in the referenced folder. During this time, a progress bar will appear that says “Inferring PMEs...”. Next, BPC uncertainties will be estimated during which time a progress bar saying, “Estimating BPC uncertainties...”, will appear. Inferring PMEs and calculating BPC uncertainties each take significant time, as discussed in the General Notes section, but *progress is saved*. If the script is interrupted, restart it in the same way as is described here, and the portions of the process that were already completed will not be run again. If a complete recalculation is desired, the files created by the program should be deleted manually. The directory structure of these saved files is described in the General Notes section.

While the script is running, MATLAB may appear not to be busy, causing confusion (see further discussion in the General Notes section). After the required calculations are complete, you should see a color-coded table output to a new figure:



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.

To re-display BPC values after the analysis is complete, simply re-run BPConeclick_GUI.m by clicking on the “Infer PMEs and display BPC” button from the main menu. Because completed analyses won’t be re-done, the function will quickly display BPC values.

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

Plotting a PME can be done only after PMEs have been inferred using ‘BPConeclick_GUI.m’ above. Click the “Plot PMEs” button on the main menu. You should see the following window:

PMEplot_GUI

Data folder Select Data Folder

Samples

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

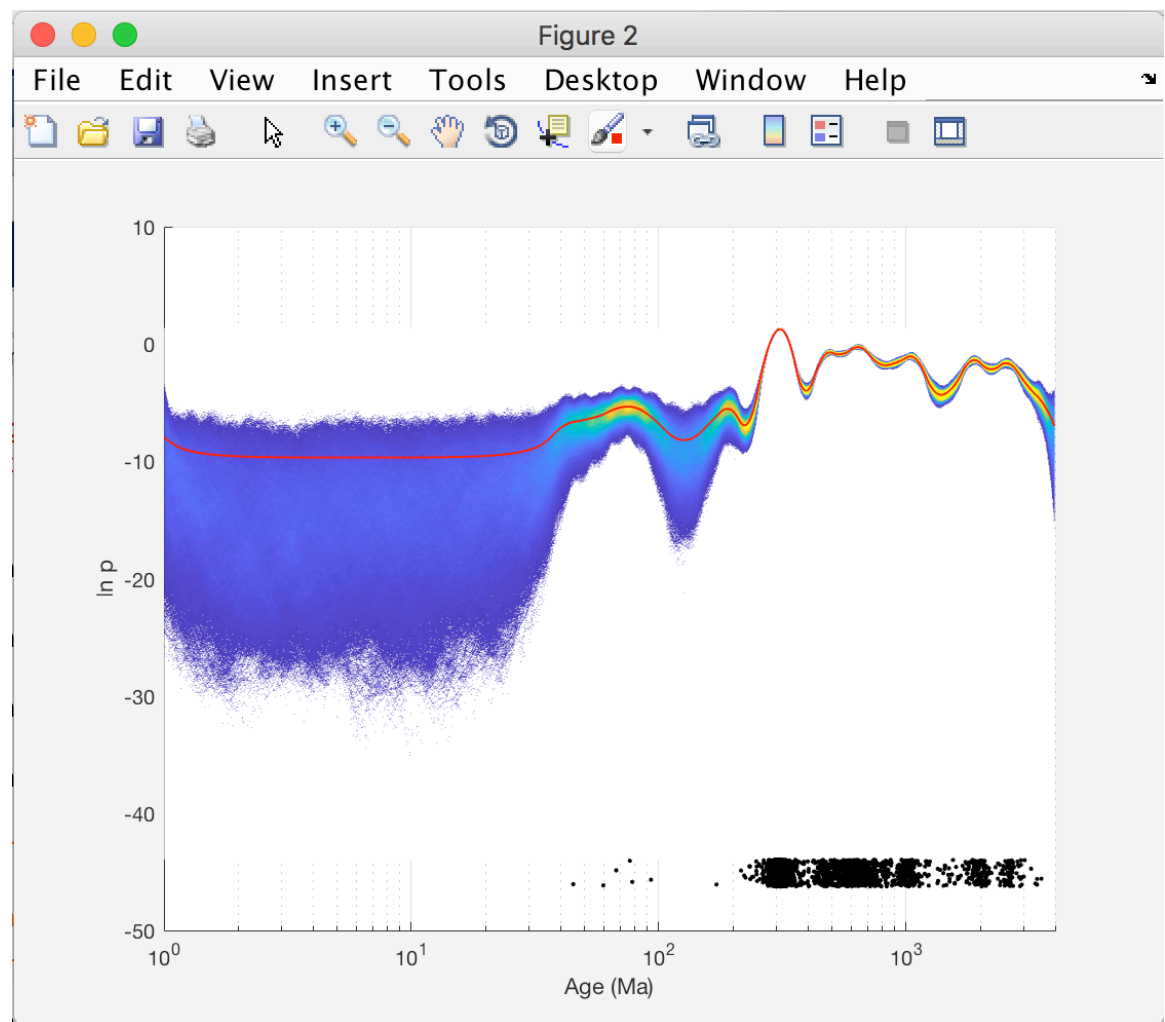
☐ Use linear probability scale (log scale is default)
☐ Show dot plot of measured ages
☐ Highlight maximum likelihood model
☐ Title figure with sample name
☐ Plot dot plot and max. likelihood model in separate figures from the PME

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

Plot PME

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. If more than one figure is plotted, then subsequent figures will be plotted in sequentially numbered plots. 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.

C. Extract probability values from PMEs using PME2CSV_GUI.m

The PDFs that constitute a PME are a collection of functions, and their values can be queried at a given set of x values using PME2CSV_GUI.m. This can be an effective way to obtain the function values of a PME for further processing. First, click “*Extract PME values to a .csv file*” in the main menu. You should see the following window:

The screenshot shows the PME2CSV_GUI window. At the top, the path is /Volumes/TYE_BLACK/dz_modelingMay16Aug2017/BPC_tosubmit_Feb2018/sample_data. A 'Select Data Folder' button is on the right. On the left, a listbox contains four items: Pullen_subsample_n300, Pullen_subsample_n60, Thomson_subsample_n300, and Thomson_subsample_n60. To the right of the listbox is a text box with instructions: 'Select the folder that contains the .csv files corresponding to sample ages and uncertainties (the same one selected for all BPC analyses). 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), and the checkbox determines whether output probability values will be logarithmic or linear. The desired number of cores may be indicated, and if the user desires a thinned subset of a PME, that is also an option. Finally, input the X values you want to query into the column of the displayed table, or import a .csv containing a single column with the desired ages values.'

Below the listbox is a table titled 'Age values (Ma) to query'. It has two columns: the first column contains numbers 1 through 8, and the second column is empty. To the right of the table are input fields for 'X min (Ma)' (value 1), 'X max (Ma)' (value 4000), a checkbox for 'Use linear probability scale (log scale is default)' (unchecked), a field for 'Thin the PME to a maximum number of models indicated here (leave blank not to thin)' (value 10000), and a field for 'Number of cores to use for analyses (leave blank to use all available)' (blank). At the bottom, there are two buttons: 'Import values from .csv file' and 'Retrieve PDF Values'.

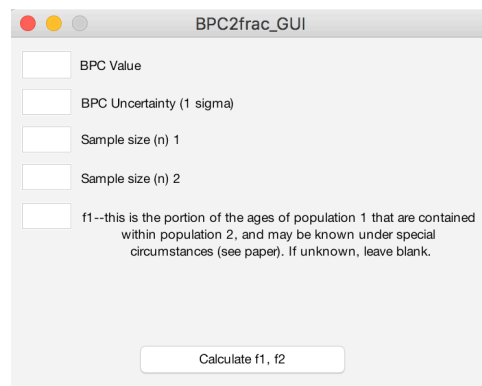
First, click “*Select Data Folder*”, which will open a directory selection dialog. The selected directory should contain the .csv files that correspond to each modeled sample—this is the same data folder that BPConeclick_GUI.m uses. The listbox will then show the names of samples with .csv files saved in the data folder. Select the sample(s) for which you want to extract values, then choose options, including the x min and max values—these must be the values that you used for modeling the data. Check the box to output linear rather than logarithmic probability values. Enter the maximum number of probability models from which you wish to extract values—a default value of 10000 is listed. Remove this default and leave the space blank to use all the probability models in a PME, but this may result in a very large output that

takes a long time to write to disk. Optionally enter a number of cores to use for processing (leave blank to use all cores available). In the lower left corner, manually enter the desired ages at which you wish to output the probability density, or load a one-column .csv file of desired ages by clicking the button. This .csv should consist of a single column with age values in Ma with no header. Then, click the “Retrieve PDF Values” button. A progress bar should appear which says “Evaluating PDFs...”, and a message box will appear when the process is complete.

The script creates a new subdirectory within the data folder you specified called ‘PMEvalues/’ and writes .csv files containing PDF values at the queried ages in that subdirectory. Output .csv files are named according to the sample name, followed by ‘PMEvals.csv’. In these output .csv files, the first column lists the queried ages. Then, each successive column displays the probability density values of one PDF at each of the queried age values. Each column (except the first) contains values from a single PDF.

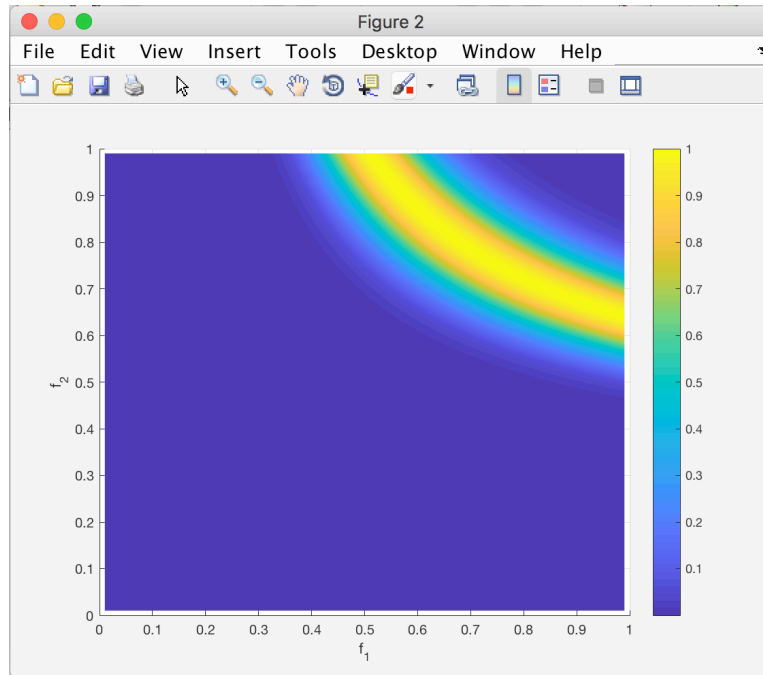
D. Inferring the shared fractions of two populations from BPC values using BPC2frac_GUI.m

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 paper 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, click the “Infer shared fractions of parent populations from BPC” button in the main menu. The following window should appear:



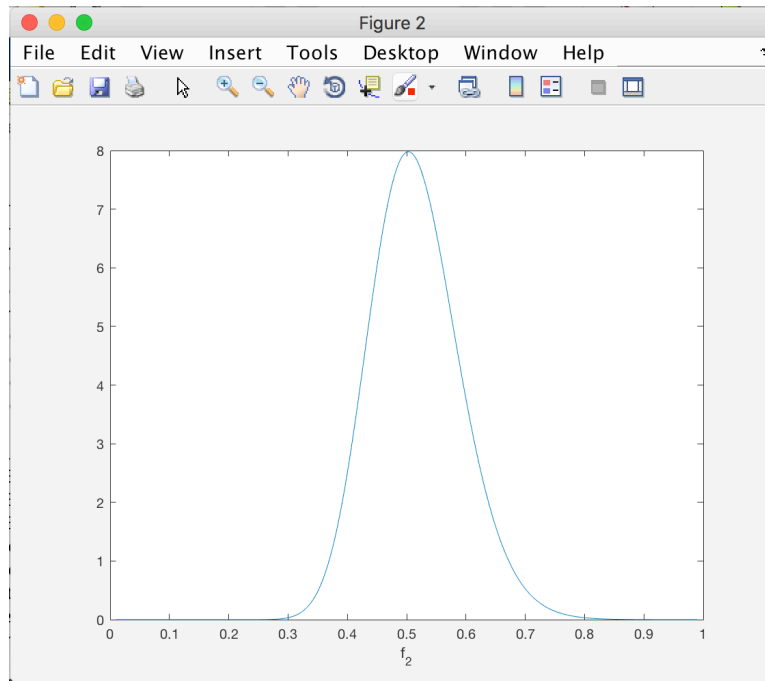
Fill out the text box fields, including BPC value and uncertainty, along with the sample sizes of the compared samples. The BPC age and uncertainty come from the results of part A. In specific circumstances, the shared proportion of one sample may be assumed. For instance, if two samples are

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 f_1 value of 1 (see text for discussion). If f_1 can be assumed, then enter it as well. If no f_1 value is entered, then the output will appear like this:



Here, colors indicate the relative likelihood of coordinate pairs of (f_1 , f_2) 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 f_1 can be assumed, the output will appear like this:

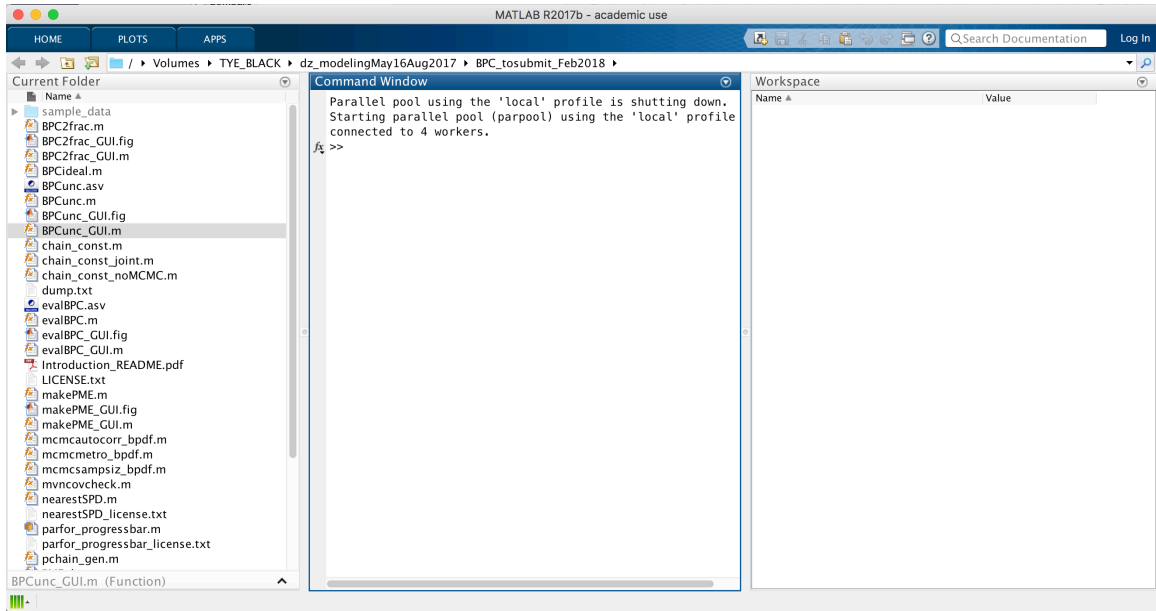


Here, the plot shows the likelihoods of different values of f_2 for the given value of f_1 . This example plot was generated using the same parameters as above, plus an f_1 value of 1. In addition, the mean and standard deviation of this distribution are output to the MATLAB Command Window.

5. General Notes, Common Problems, Clarifications

MATLAB may appear not to be busy even though the script is running: Once you start any of these scripts from the GUI (by clicking “Generate PMEs” or equivalent), the progress bar may appear under other MATLAB windows. In addition, a blinking cursor will appear in the MATLAB Command Window, and MATLAB will not display the word “Busy”, as it usually does for a script run from the command line, even though the script is running. The screenshot below shows the MATLAB screen while the script is running. **This behavior will occur with all scripts included in our package when run from GUIs.**

The standalone applications can be slow to respond: As noted above, when the standalone applications are started or a button is clicked, the response of the application may be delayed. The standalone application will often be far more delayed than the MATLAB script. When this occurs, please wait for the application to respond rather than reissuing a command, which will cause operations to be run more than once.



Expected time for script running: The scripts `makePME.m` and `BPCunc.m` (and equivalently `makePME_GUI.m` and `BPCunc_GUI.m`) will take a significant amount of time to run. These scripts compare every possible pair of samples in an input dataset. Mathematically, the number of possible pairs in a given dataset is given by $(N \text{ choose } 2)$, where N is the number of samples in a given dataset. `makePME.m` and `BPCunc.m` both take minutes to tens of minutes for each sample plus an equivalent time for each possible sample pair. This time is distributed over as many cores as are available for processing, such that the total processing time (in hours) can be estimated as

$$T = (0.2 - 0.5 \text{ hours}) * \frac{N + (N \text{ choose } 2)}{\# \text{ cores}}$$

These times were calculated on a 2014 Macbook Pro with 2.2 GHz, 4-core processor and 16 GB RAM.

Progress is saved when running the scripts. The scripts that run for significant lengths of time (`makePME.m` and `BPCunc.m`, or equivalently `makePME_GUI.m` and `BPCunc_GUI.m`) run a Monte Carlo-based analysis on each sample and each possible pair of samples in the input dataset, which take approximately minutes to tens of minutes per sample or sample pair, as discussed in the above note. These scripts immediately save the results of their analysis on each sample or pair of samples as soon as the analysis is complete. The scripts look for these files when run and do not run duplicate analyses for samples or sample pairs that have already been completed. Thus, if one of these scripts is interrupted in the middle of execution, it can be restarted without losing the analyses that have already been completed. Effectively, your progress is saved as the script runs. However, this feature also means that if one of the files that the scripts write and read becomes

corrupted or somehow unusable to the script, it must be deleted manually in order to signal to the script to redo that analysis. If you suspect that such a corruption has occurred, for ease I suggest deleting the 'chains/', 'log/', and 'unc/' folders created by the scripts and re-running them. See the section "Workflow Detail" about the relevant scripts for a description of the file architecture.

File architecture of makePME.m and associated scripts for PME

inference: 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".

File architecture of BPCunc.m and associated scripts for estimating BPC

uncertainties: BPCunc.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.

An additional subdirectory, 'unc/log/', is created, where a log file is stored for each sample or pair of samples for which BPC uncertainty is estimated.

Age bounds used for modeling PME are not saved. 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.

6. 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.