

DZnmf2D v.1.0

User Guide

Joel E. Saylor¹

Kurt E. Sundell²

1: Department of Earth, Oceans and Atmospheric Sciences, University of British Columbia,
Vancouver, BC, Canada

2: Department of Geosciences, University of Arizona, Tucson, AZ, USA

Contents

1. Introduction	3
2. Setup	3
3. Accessing the software	3
4. Data input format	5
4.1. Input one-dimensional data sets	5
4.2. Input two-dimensional data sets	5
5. Loading data	7
6. Age distributions	8
6.1. One-dimensional distribution options	8
6.2. Two-dimensional distribution options	8
7. Activating, Merging, and Deactivating Samples	10
8. Stopping criteria	13
8.1. Default one-dimensional stopping criteria	13
8.2. Default two-dimensional stopping criteria	13
9. Run NMF	14
10. Determining the optimum number of sources	16
11. Forcing multiple factorization runs	18
12. Saving figures and data	19
13. Comparing factorized and empirical sources	20

1. Introduction

DZnmf2D is a MATLAB-based graphical user interface (GUI) designed to implement non-negative matrix factorization (NMF) of univariate or bivariate detrital provenance data sets. It builds on DZnmf2D software developed by Saylor et al. (2019). The program can be run in MATLAB, or as a stand-alone GUI in Windows or macOS.

2. Setup

Running DZnmf2D v.1.0 outside of a MATLAB environment requires installation of the MATLAB Compiler Runtime (MCR) version R2020a (9.8). The compiler allows the DZnmf2D GUI to run on the user's computer with installing MATLAB.

Download the MCR here: <https://www.mathworks.com/products/compiler/matlab-runtime.html>

Select all the default settings and install (this takes some time, usually about 15 minutes). The MCR will create a new directory on the user's machine that will be accessed in the background when the GUI is running; it will not interact with any existing versions of MATLAB or other versions of the MCR on the user's machine.

The DZnmf2D GUI does not require any installation other than the MCR compiler. Once the MCR is installed, download the DZnmf2D GUI and example data sets from the Supplemental Material in Saylor et al. (in review), <https://www.kurtsundell.com/>, <https://www.joelesaylor.com/>, or <https://github.com/kurtsundell>.

3. Accessing the software

The software can be saved anywhere on the user's computer. Once the software is downloaded and the MCR installed, the software can be accessed by selecting the GUI file (i.e., DZnmf2D_v001.exe) in the folder where it was saved. As long as the default settings were selected for the MCR it will be accessible by the DZnmf2D GUI. Occasionally there is an error in the MCR installation, which will throw the error message in Figure 1.

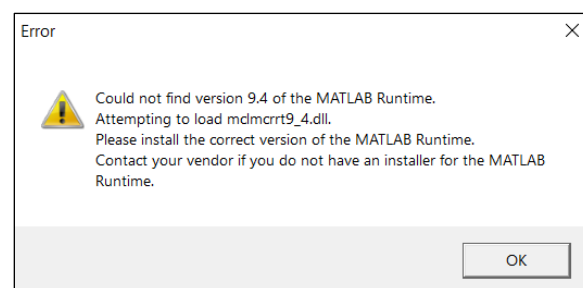


Figure 1. Error message generated when the GUI cannot access the MATLAB Compiler Runtime (MCR).

If the error message in Figure 1 or a similar error message is thrown, it means the DZnmf2D GUI cannot find and/or access the MCR. If this happens there are two possible solutions to troubleshoot: (1) restart the machine and try again, or (2) reinstall the MCR (the MCR may need to be completely removed and directory deleted before reinstalling rather than simply overwriting). If these options do not work please contact one of the first two authors.

Upon opening the DZnmf2D GUI there will be two blank listboxes and two blank plots (Figure 2). All of the program functionality (i.e., analysis, plotting, exporting, etc.) is contained within this single main page.

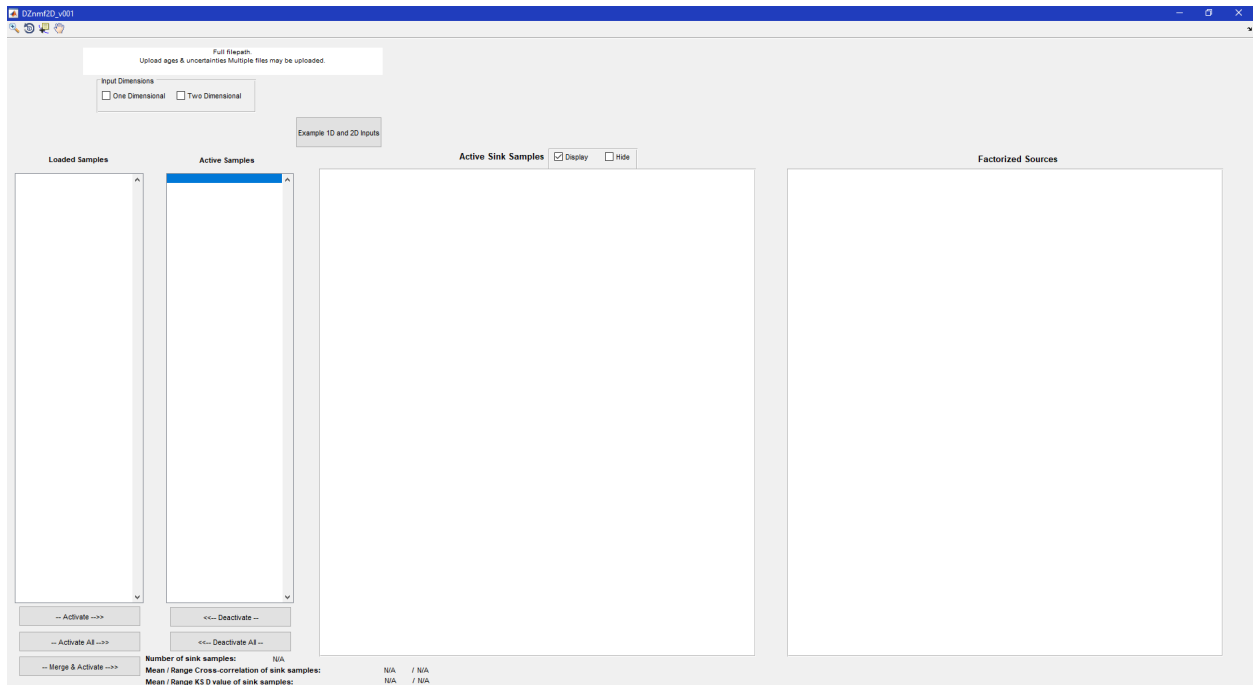


Figure 2. Initial DZnmf2D window.

4. Data input format

Example input files can be accessed by selecting the “Example 1D and 2D inputs” (Figure 3). Input data is formatted so each sample has two columns.

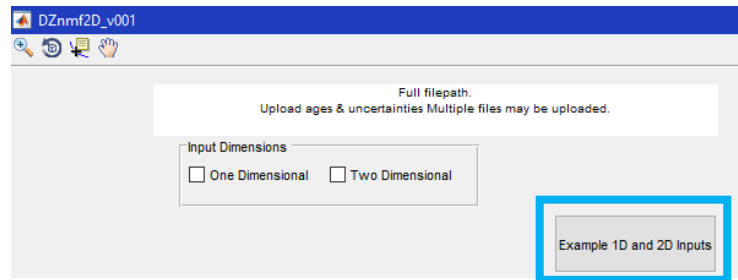


Figure 3. Example 1D and 2D Inputs button indicated by the blue box.

4.1. Input one-dimensional data sets

For univariate distributions, these two columns are mean age and age uncertainty (Figure 4).

Tapeats 1	1 Sigma	Tapeats 2	1 Sigma	Lower Bright...	1 Sigma
1.4533e+03	51.5808	1.6778e+03	44.0351	1.7048e+03	10.0811
1.0062e+03	71.8513	1.6661e+03	32.1235	1.7548e+03	23.8771
1.4882e+03	34.1113	1.7203e+03	47.8193	1.7149e+03	37.4208
1.7284e+03	43.5099	1.4425e+03	48.4682	1.4542e+03	12.0389
1.4796e+03	44.7515	1.2037e+03	54.1104	1.7623e+03	12.8853
1.5646e+03	20.9944	1.4474e+03	23.5206	1.7662e+03	24.0167
1.7236e+03	51.2609	1.6495e+03	42.6159	1.7416e+03	11.6051
1.7047e+03	34.4400	1.7522e+03	35.8784	1.6994e+03	24.3069
1.6705e+03	20.3519	1.7702e+03	40.0112	1.7047e+03	23.7010
1.4689e+03	20.9015	1.7811e+03	44.1040	1.7030e+03	25.3982
1.4605e+03	22.1070	1.4364e+03	34.9916	1.7745e+03	22.6951
1.6976e+03	29.5131	1.7702e+03	38.9224	1.4396e+03	28.3054
1.7056e+03	44.2260	2.6846e+03	29.1123	1.7855e+03	8.9534
1.6951e+03	39.4544	1.7501e+03	57.1565	1.7500e+03	16.5948
1.6628e+03	25.9542	1.4593e+03	33.1659	1.9299e+03	17.6446
1.7965e+03	40.7765	1.6525e+03	48.7125	1.7715e+03	26.3039

Data from Gehrels, G. E., Blakey, R., Karlstrom, K. E., Timmons, J. M., Dickinson, B., & Pecha, M. (2011). Detrital zircon U-Pb geochronology of Paleozoic strata in the Grand Canyon, Arizona. Lithosphere, 3(3), 183-200.

Figure 4. Example 1D data sets, accessible by selecting the “Example 1D and 2D Inputs” button (Figure 3).

4.2. Input two-dimensional data sets

For bivariate distributions, these two columns are the mean of the first and second variables (Age and ϵHf_T value in Figure 5).

Example 2D data set for DZnmf2D (.xls, .xlsx, .csv, etc.)

S1 Age	S1 Ehft	S2 age	S2 Ehft	S3 age	S3 Ehft
3.1417e+03	0.8180	599.1000	8.1720	2136	1.6690
2151	2.4980	703	5.7040	759	1.6700
535	-0.9400	1226	-1.1550	2671	-10.7400
622	-1.0130	697.5000	7.3410	3522	1.5780
2095	2.5600	557	5.1930	1111	3.3380
3195	-0.2850	796	-0.1320	1867	-4.5020
1113	3.8830	596	-5.7970	3.0887e+03	0.3770
735.8000	-9.5790	660	9.4730	894	12.1000
514	-9.0110	627	-9.4420	2.5896e+03	-7.1500
2121	6.0580	1810	-3.2140	3338	-0.5420
247	-11.2540	3066	-6.5110	696	2.9040
1.0886e+03	11.2280	980	-0.4510	3241	-5.4770
673	-1.6550	1118	-15.7720	2174	2.7220
2721	0.4750	607	-3.9870	3.1137e+03	-0.0660
617	6.8150	1531	-9.6660	910	6.7120
3230	-0.2390	941	0.1770	610.8000	-11.5970
2.6633e+03	-7.3430	1189	-0.5950	296	1.5000

Data from Puetz, S. J., and K. C. Condie (2019), Time series analysis of mantle cycles Part I: Periodicities and correlations among seven global isotopic databases, *Geoscience Frontiers*, 10(4), 1305-1326.

Figure 5. Example 2D data sets, accessible by selecting the “Example 1D and 2D Inputs” button (Figure 3).

5. Loading data

To load data, first select whether the data to be loaded are 1D or 2D distributions by selecting the “One Dimensional” or “Two Dimensional” checkbox (Figure 6). If the data to be loaded is One Dimensional, also select the uncertainty level.

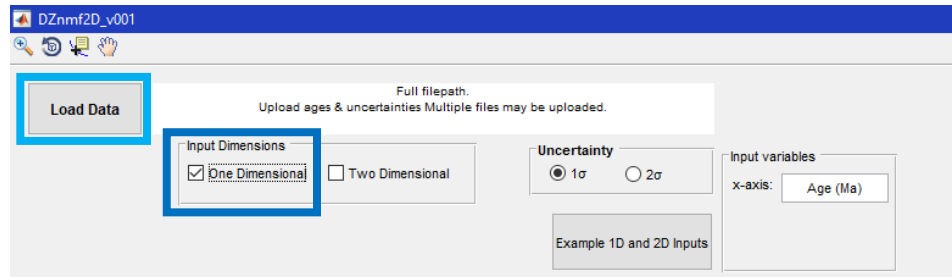


Figure 6. To load data, first indicate whether the data to be loaded is One Dimensional or Two Dimensional distributions (dark blue box). Then select “Load Data” (light blue box) to open a browser and navigate to the input file location.

Select “Load Data” (Figure 6) to open a browser and select the input file. Once the file is loaded, the samples in the file will populate the “Loaded Samples” window (Figure 7).

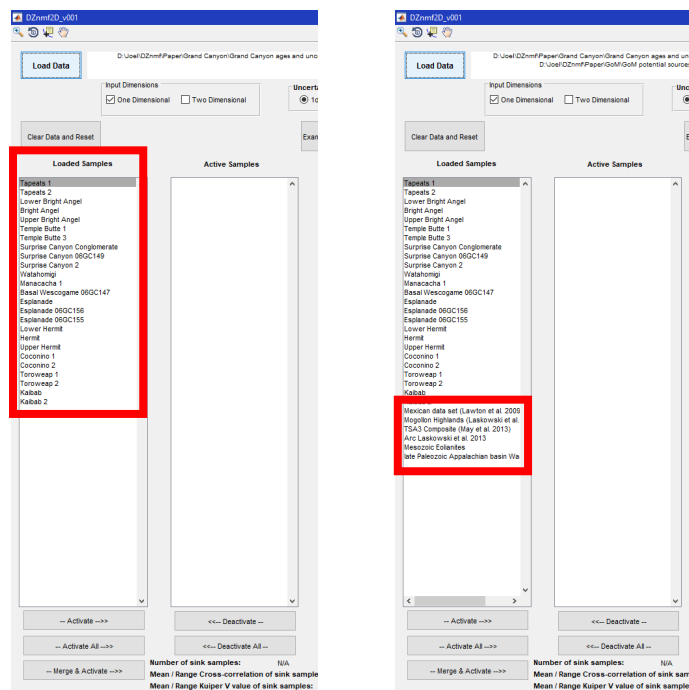


Figure 7. DZnmf2D main screen with samples loaded shown in the “Loaded Samples” listbox. Data files can be loaded one at a time. The Loaded Samples listbox shows the data sample names (left). Multiple files can be uploaded one at a time, the figure on the right shows a second input appended to the existing Grand Canyon data.

6. Age distributions

The next step is to specify how the distributions will be made. There are options to make probability density plots (PDPs) or kernel density estimates (KDEs) (Figure 6). This needs to be specified before activating the samples. Once samples are activated these options will be grayed out to ensure all distributions are generated the same way.

6.1. One-dimensional distribution options

If the loaded data are One Dimensional distributions the user can input axes titles in the “Input Variables” textbox(es) (Figure 8).

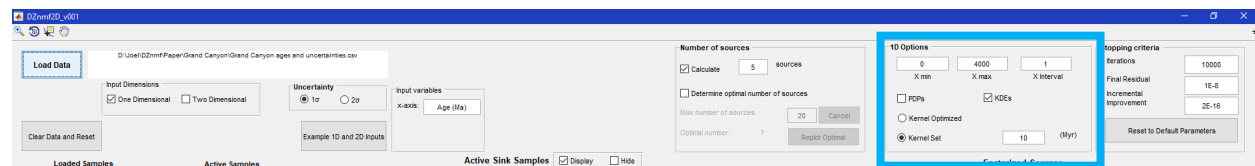


Figure 8. DZnmf2D with one dimensional distributions loaded. The blue box shows the 1D distribution options.

The user can display and factorize the distributions as Probability Density Plots (PDPs) or Kernel Density Estimates (KDEs) in the 1D Options box (Figure 8). If KDE is selected, the uncertainty in the input file is ignored and kernel bandwidth is based either on an optimized bandwidth (Botev et al., 2010) or constant bandwidth (Figure 8). The user can also specify the minimum and maximum values for the distribution and the discretization interval (X min, X max, and X interval, respectively).

6.2. Two-dimensional distribution options

If the loaded data are two dimensional distributions the user can input axes titles for both x and y axes in the “Input Variables” textbox(es) (Figure 8).

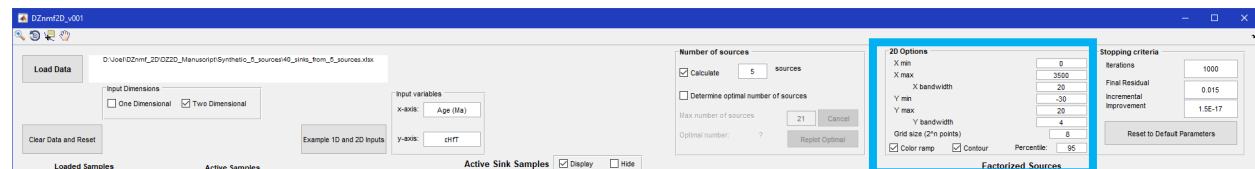


Figure 9. DZnmf2D with two dimensional distributions loaded. The blue box shows the 2D distribution options.

Two dimensional distributions are displayed and factorized as 2 dimensional KDEs with fixed bandwidths in the x and y directions. Distribution parameters can be modified in the 2D Options box (Figure 9). Parameters include the minimum and maximum x and y values, the x and y bandwidths, and the resolution of the discretization of the bivariate KDE. The user can also specify whether the distribution is plotted as an intensity map, intensity map with contours, or a clipped contour plot (Figure 10).

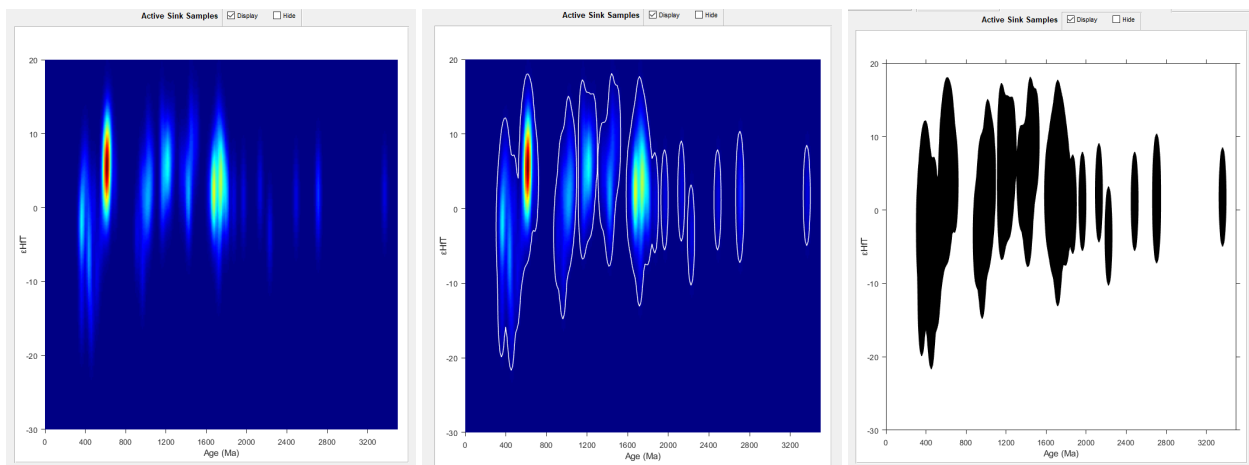


Figure 10. A bivariate data set plotted as an intensity map (left), intensity map with 95th percentile contours (middle), or clipped 95th percentile contour plot (left).

7. Activating, Merging, and Deactivating Samples

Next, select the samples from the **Loaded Samples** listbox and press the **Activate** pushbutton. Alternatively, all samples may be activated by pressing the **Activate All** pushbutton (Figure 11).

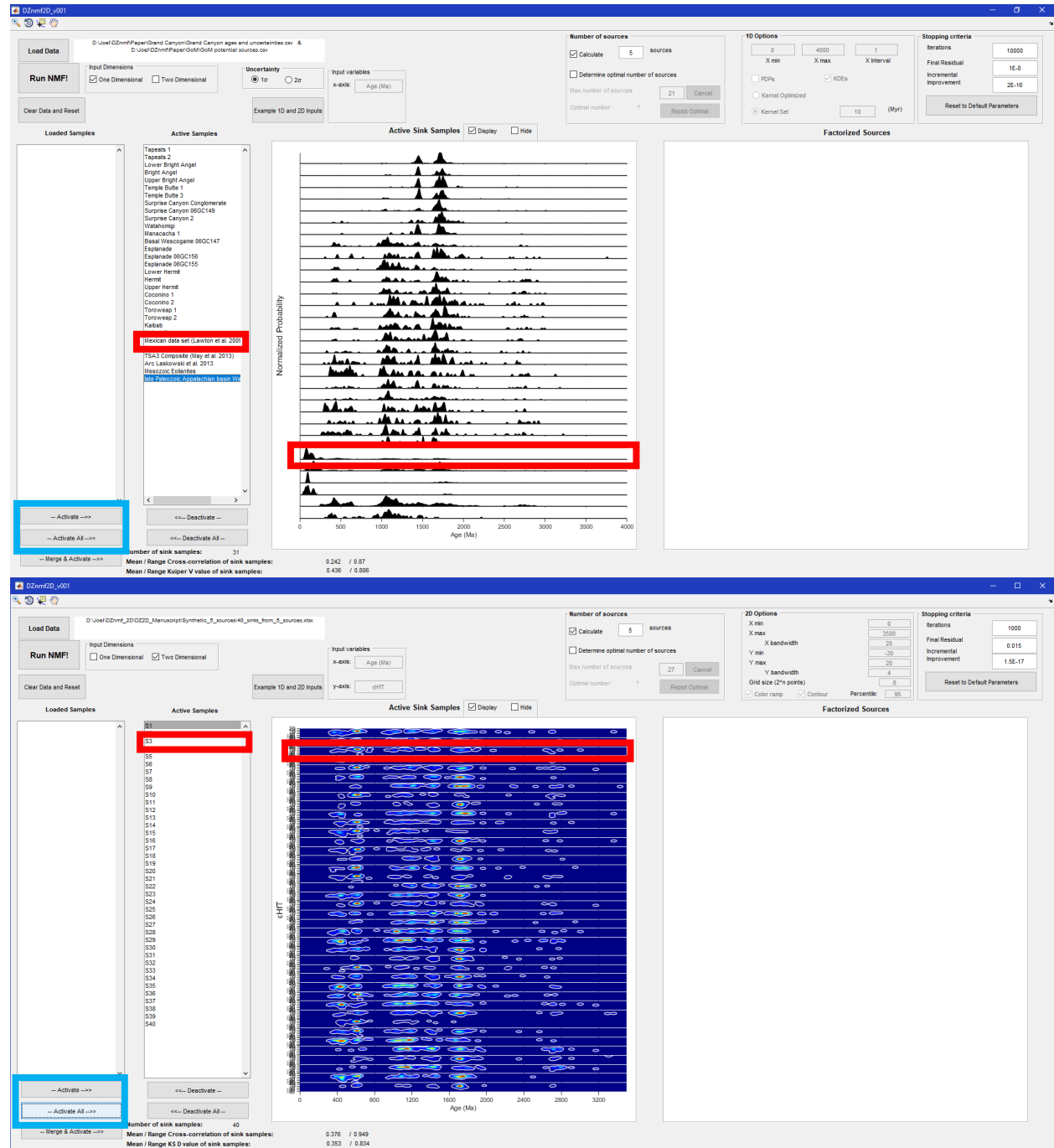


Figure 11. Activate samples by selecting the loaded samples and pressing Activate or Activate All (Blue box). Samples are arranged in the Active Sink Samples window in the same order as the Active Samples listbox. The red boxes highlight equivalent samples. Top: One-dimensional distributions. Bottom: Two-dimensional distributions

The user has the option to merge (combine) data sets (Figure 12). When samples are merged their sample names will be horizontally concatenated (Figure 12). This is the same as combining ages and uncertainties into a single age distribution. The new merged distribution will be added to the **Active Sink Samples** plot on the bottom. Note the plotted distributions are always plotted in the same order as the **Active Samples** listbox (Figure 12).

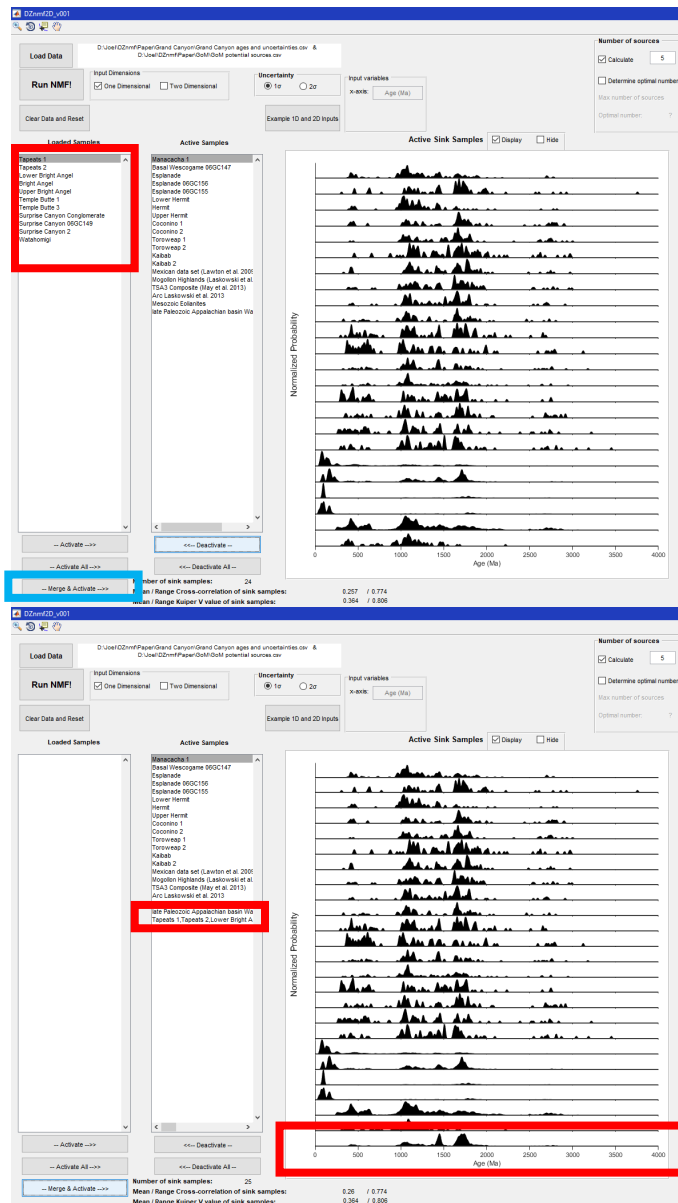
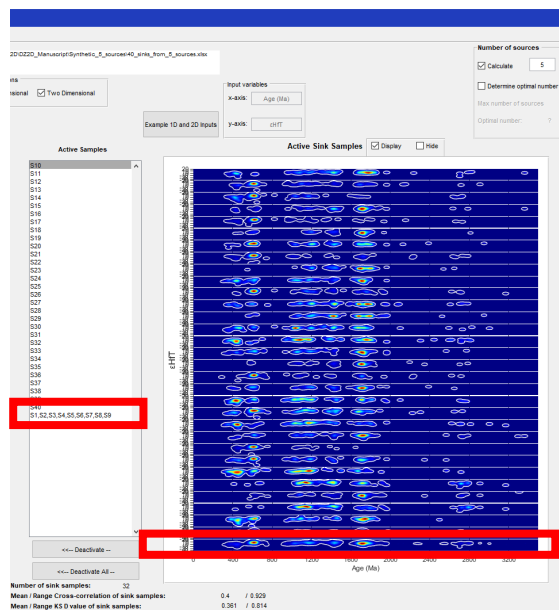


Figure 12. Merging samples. Top left: Select samples to merge. Below and Bottom left: Merged sample is added to the Active Sink Samples plot, which is always in the same order as the Active Samples listbox.



Active samples may be removed by highlighting individual sample names and pressing the **Deactivate** pushbutton (Figure 13). The user also has the option to **Deactivate All**. This will clear the **Active Sink Samples** plot and return all of the samples to the **Loaded Samples** listbox in their original order, and the distribution options (Figures 8 and 9) will be made available.

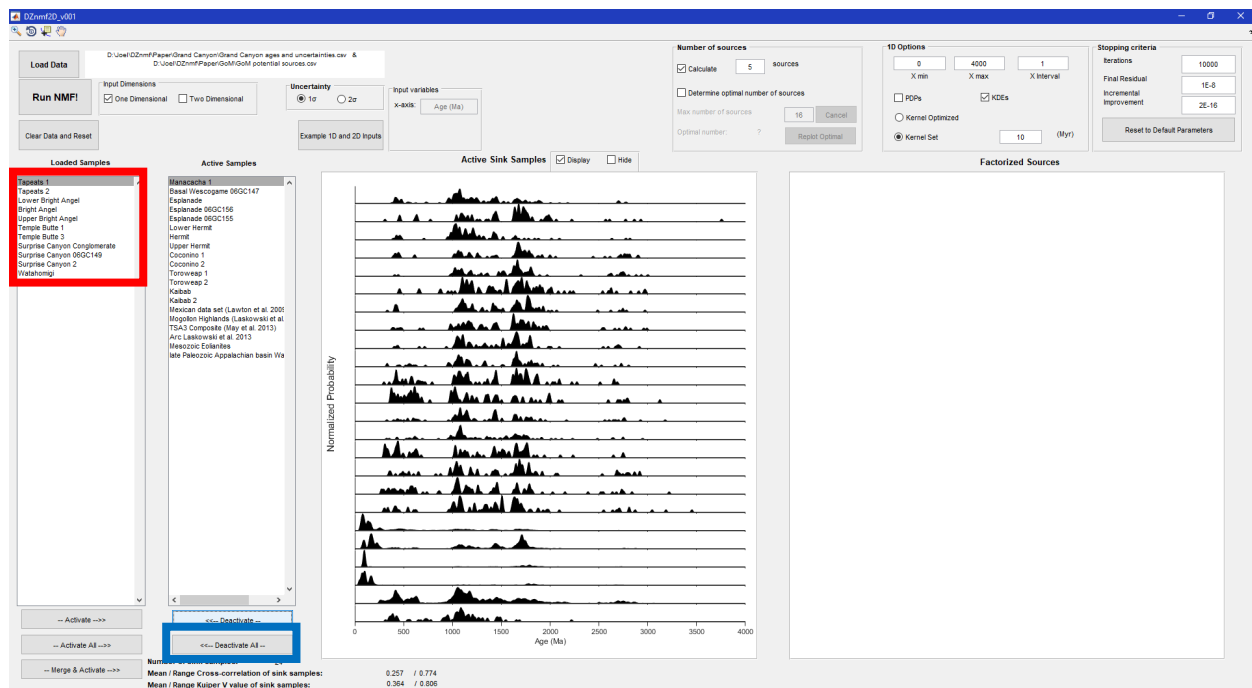
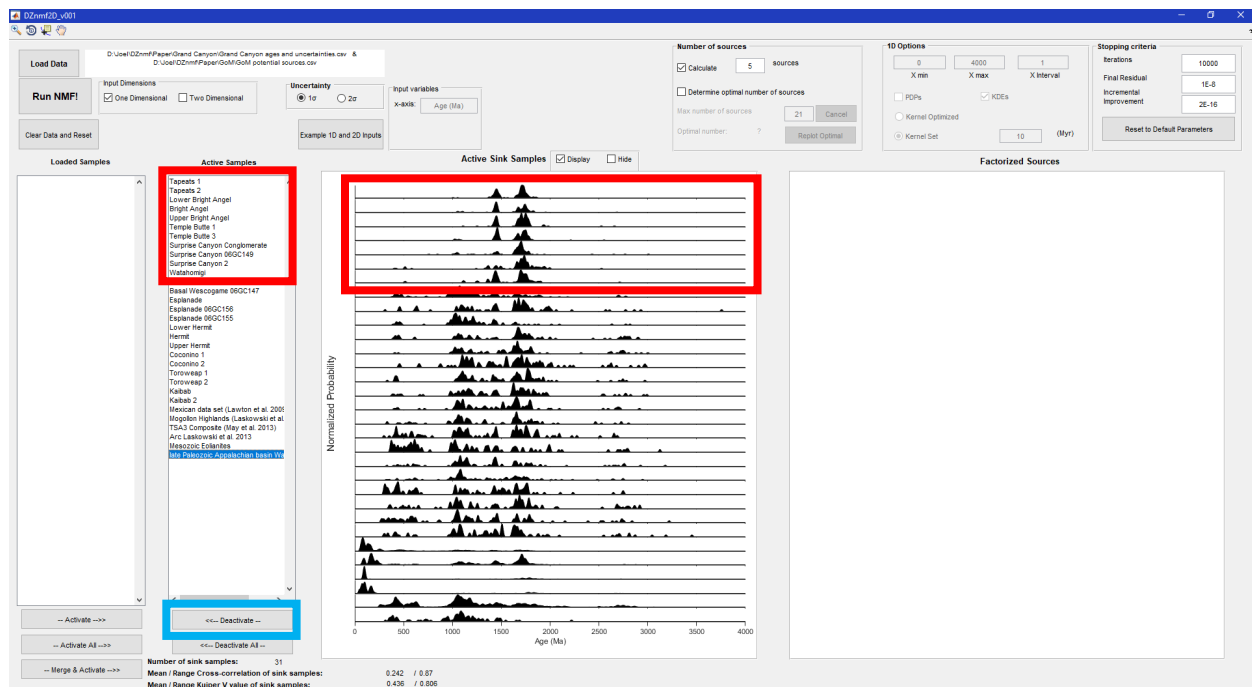


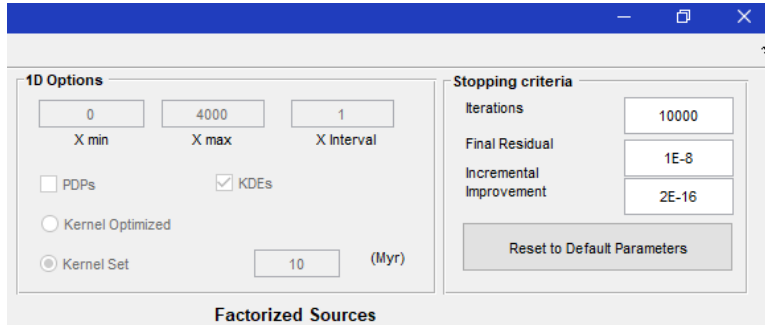
Figure 13. Samples can be deactivated by highlighting the samples in the Active Samples listbox and selecting Deactivate (light blue box, top panel). Alternatively, all samples can be deactivated by selecting Deactivate All (dark blue box, lower panel)

8. Stopping criteria

The default stopping criteria are suitable for most applications to detrital geochronology, and will rarely need to be changed. However, the user has the option to change these parameters and to reset to default parameters (Figure 14). In particular, the stopping criteria can be changed to force the algorithm to run multiple factorizations (see section 11).

8.1. Default one-dimensional stopping criteria

The default stopping criteria is set to 10,000 iterations, $1\text{E-}8$ final residual, and $2\text{E-}16$ incremental improvement (Figure 14).



The screenshot shows a software window titled "1D Options". It contains two main sections: "1D Options" and "Stopping criteria". In the "1D Options" section, there are three input fields: "X min" (0), "X max" (4000), and "X Interval" (1). Below these are checkboxes for "PDPs" (unchecked), "KDEs" (checked), "Kernel Optimized" (radio button, unselected), and "Kernel Set" (radio button, selected). There is also a "10 (Myr)" field. The "Stopping criteria" section has three input fields: "Iterations" (10000), "Final Residual" ($1\text{E-}8$), and "Incremental Improvement" ($2\text{E-}16$). A "Reset to Default Parameters" button is at the bottom. The window title bar has standard minimize, maximize, and close buttons.

Figure 14. Default stopping criteria for a one-dimensional NMF run. Note that the 1D Options have been greyed out in this figure because the distributions have been activated.

8.2. Default two-dimensional stopping criteria

The default stopping criteria is set to 1,000 iterations, 0.015 final residual, and $1.5\text{E-}17$ incremental improvement (Figure 15).



The screenshot shows a software window titled "2D Options". It contains two main sections: "2D Options" and "Stopping criteria". In the "2D Options" section, there are several input fields: "X min" (0), "X max" (3500), "X bandwidth" (20), "Y min" (-30), "Y max" (20), "Y bandwidth" (4), and "Grid size (2*n points)" (8). There are also checkboxes for "Color ramp" (checked) and "Contour" (checked), and a "Percentile:" field (95). The "Stopping criteria" section has three input fields: "Iterations" (1000), "Final Residual" (0.015), and "Incremental Improvement" ($1.5\text{E-}17$). A "Reset to Default Parameters" button is at the bottom. The window title bar has standard minimize, maximize, and close buttons.

Figure 15. Default stopping criteria for a two-dimensional NMF run. Note that the 1D Options have been greyed out in this figure because the distributions have been activated.

9. Run NMF

Once the data are loaded and samples are activated the NMF may be run. The default setting is a specified number source samples to reconstruct (factorize) from the active sink samples. The user has the option to select how many sources to factorize (Figure 16). Press the **RUN NMF!** pushbutton to run NMF.

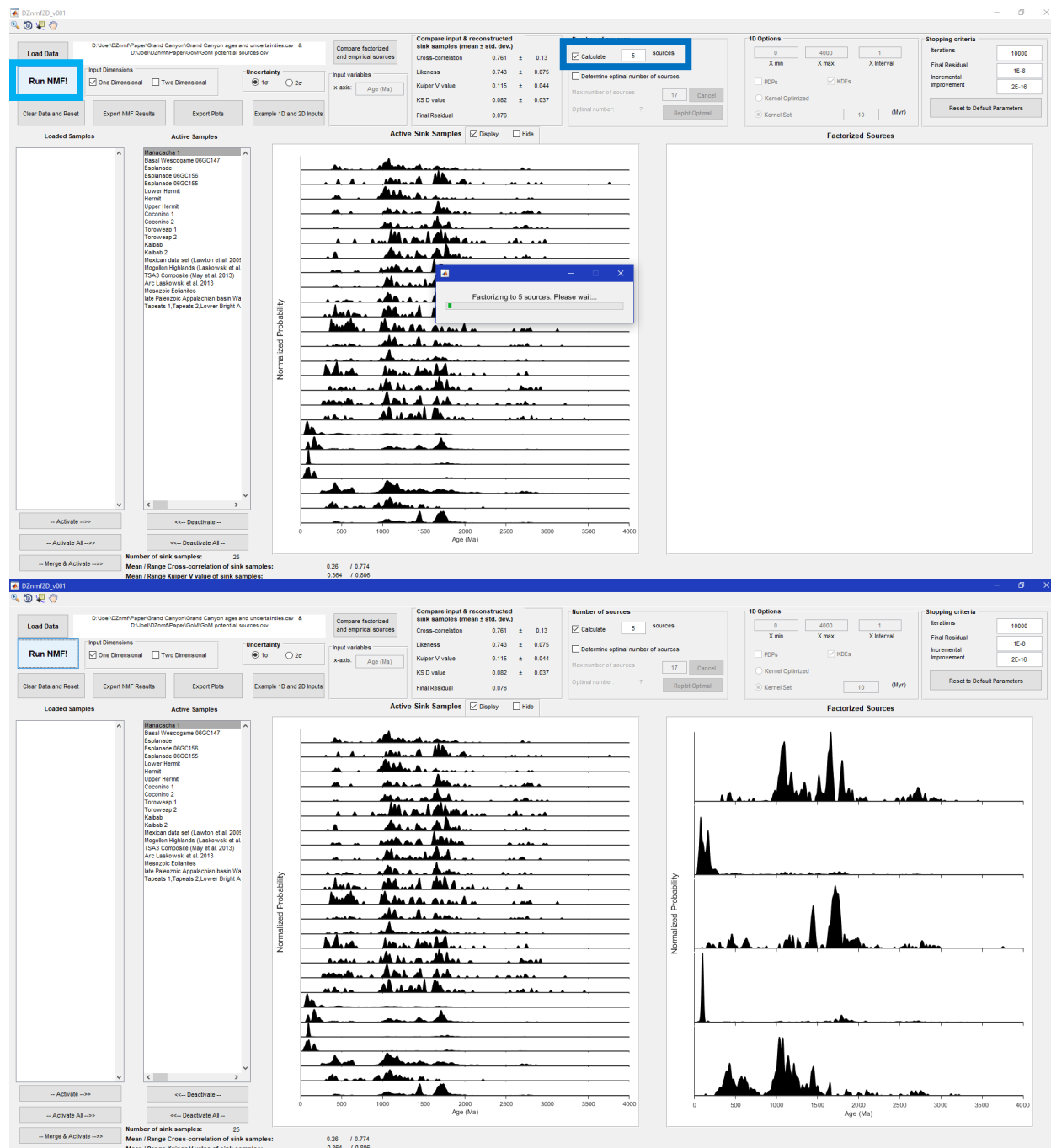


Figure 16. Factorization of five sources. Upper: Select five sources to from Number of sources panel and push Run NMF! pushbutton. Lower: Five factorized sources

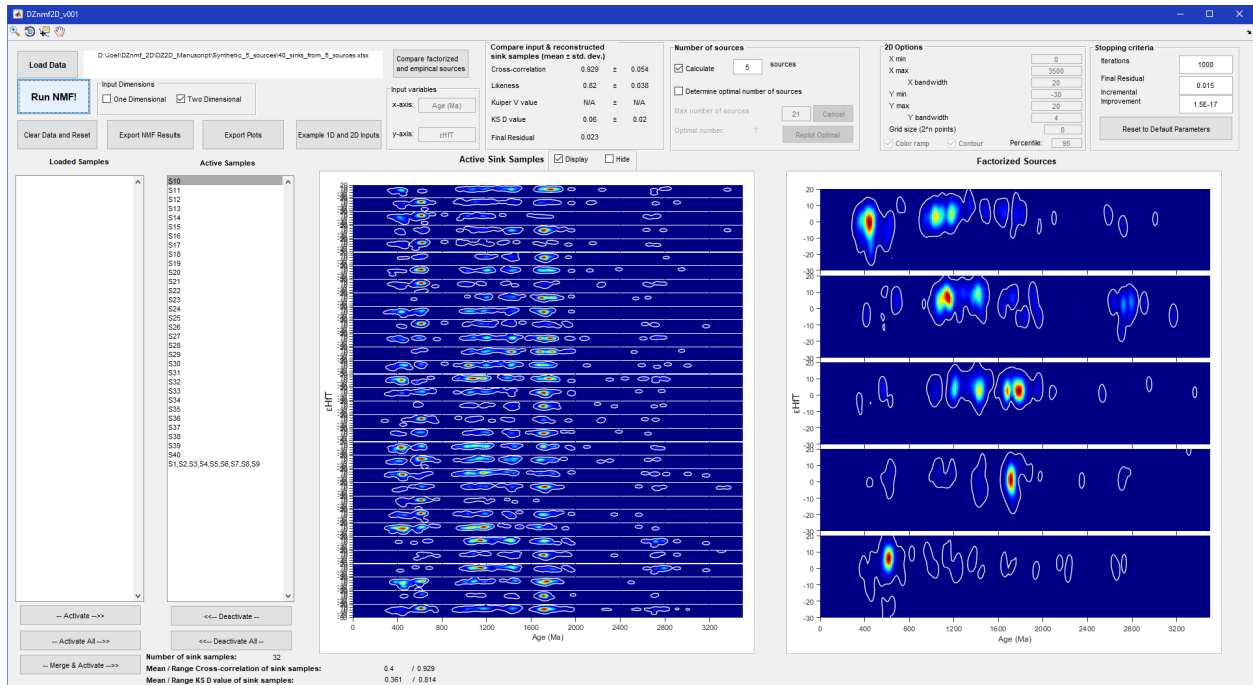


Figure 17. Factorization of a two-dimensional data set proceeds as for a univariate data set.

10. Determining the optimum number of sources

Instead of specifying the number of sources to factorize, an alternative approach is to specify the maximum number of sources (N), and determine the optimal number of sources to factorize. Here, the program will individually factorize to 2 through N sources, and plot the optimal number of sources based on calculating the final residual between input and reconstructed sink samples over the range of ranks (2 to N); see section 3.2 of Saylor et al. (2019) for a detailed description. The user has the option to cancel the run (Figure 18).

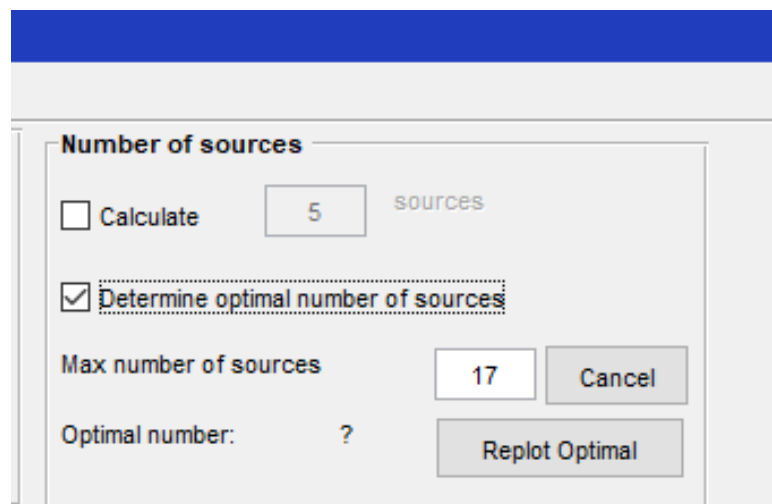


Figure 18. Check box for determining optimal number of sources. Edit text box specifies the maximum number of sources to factorize. The default Max number of sources is 2/3 of the total number of sink samples.

Once DZnmf2D has factorized the data set to all ranks between 2 and 17 it will calculate the sum of squared residuals for two linear segments fit to the final residuals (Figure 19). The optimum rank is then that rank that minimizes the sum of squared residuals for the two linear segments (Figure 19).

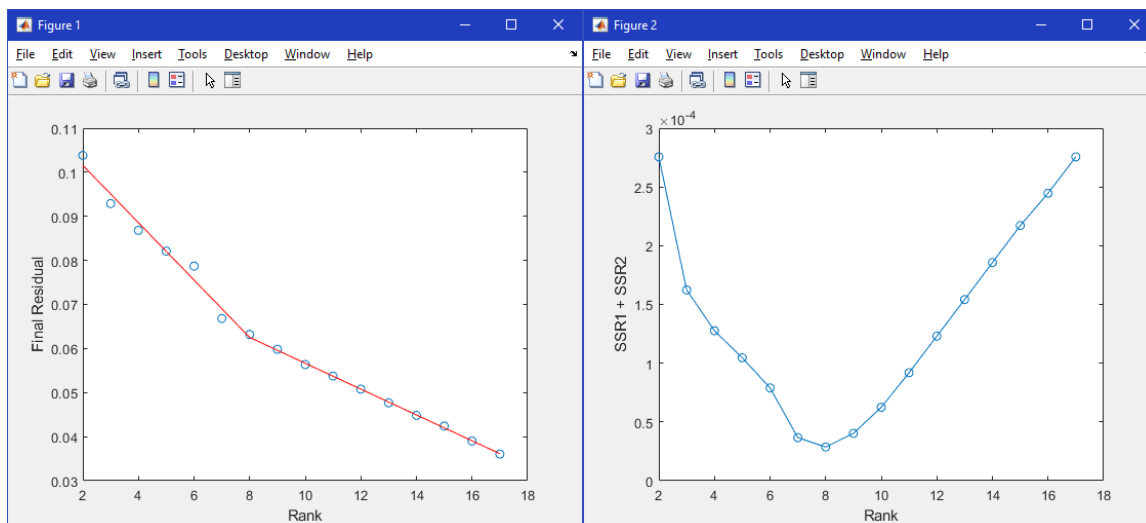


Figure 19. Left pane. Final residuals for ranks 2–17 plotted as open circles and the segmented linear regression and minimizes the sum of squared residuals plotted as a red line. Right panel. A minimum in the sum of squared residuals for the linear regression indicates the optimum rank (in this case 8).

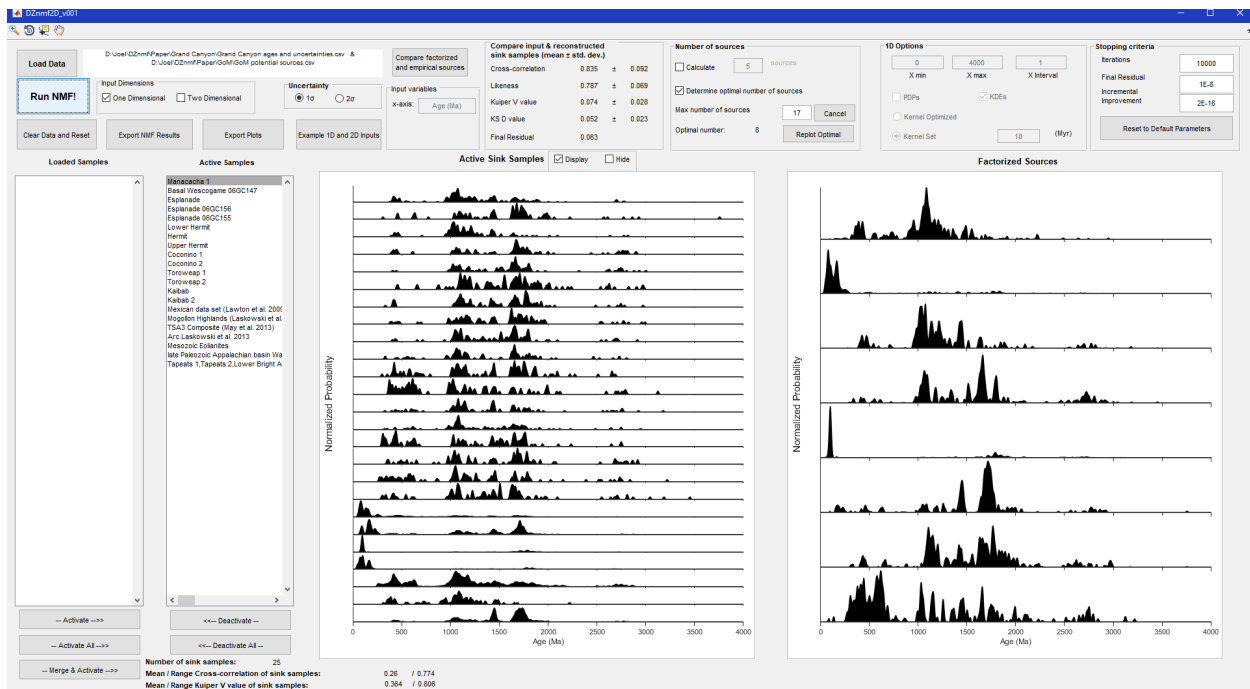


Figure 20. After factorizing all ranks from 2–17, the software will plot the factorization with the lowest final residual (8 in this case).

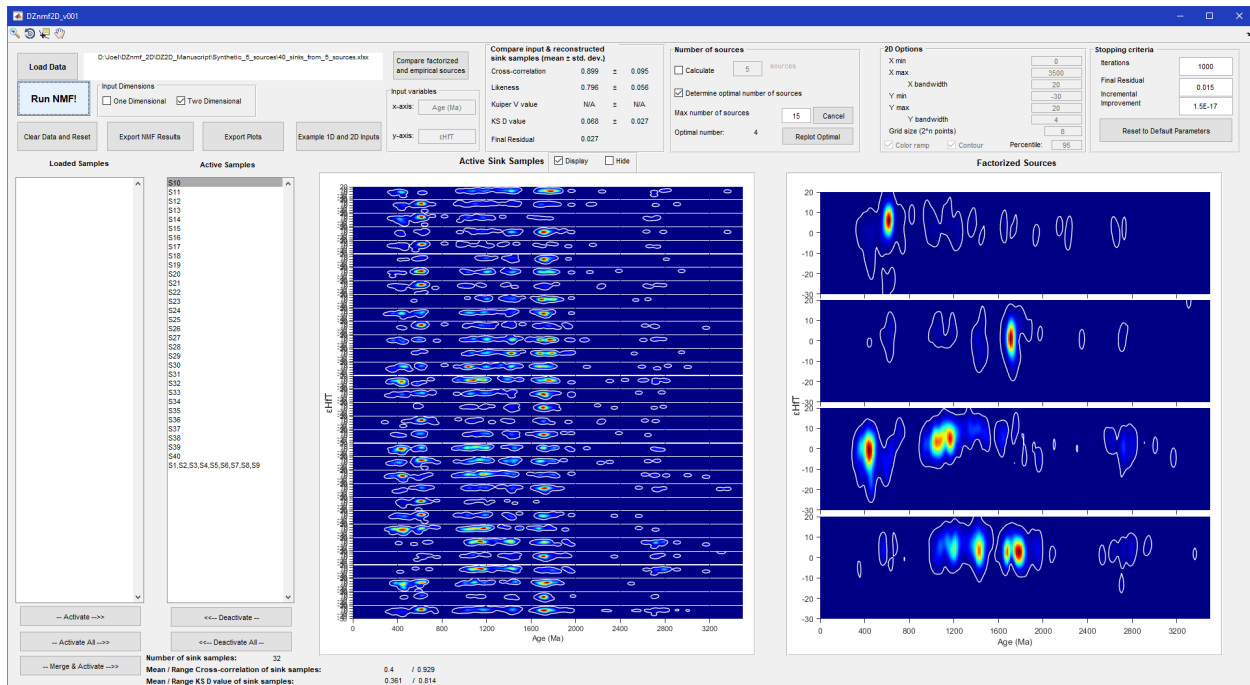


Figure 21. After factorizing all ranks from 2–15, the software will plot the factorization with the lowest final residual (4 in this case).

11. Forcing multiple factorization runs

The NMF model may result in nonunique solutions. This typically happens if there are a large number of sources, distributions may be factorized into individual components of those distributions (typically due to sediment recycling), or if the active samples are very similar. Hence we recommend that once the optimal number of sources is found, the user rerun NMF at that rank. In this case, if the NMF model does not converge on a solution within the specific number of iterations it will ask the user if they would like to repeat the calculation starting from a different random starting point in order to find the best solution.

The algorithm can be forced to run multiple iterations by setting the Final Residual and Incremental Improvement to arbitrarily low values ($<1E-100$, Figure 22). Because the Final Residual or Incremental Improvement will not be reached it will force the software to ask if the user wants to run multiple iterations (Figure 23). Once the software has completed the requested number of iterations it will retain the factorization with the lowest final residual.

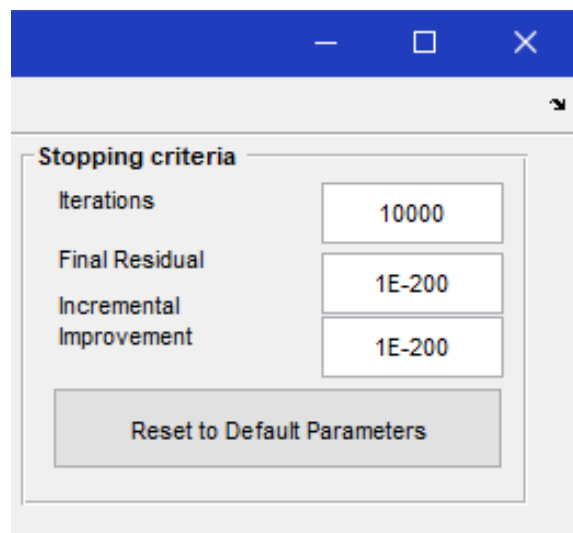
A screenshot of a software window titled "Stopping criteria". It contains three input fields: "Iterations" with the value "10000", "Final Residual" with the value "1E-200", and "Incremental Improvement" with the value "1E-200". Below these fields is a button labeled "Reset to Default Parameters".

Figure 22. Setting the stopping criteria to arbitrarily low values will force the software to ask whether the user was to run multiple iterations. This approach can be used to search the factorization space to identify a globally optimized (rather than locally optimized) result.

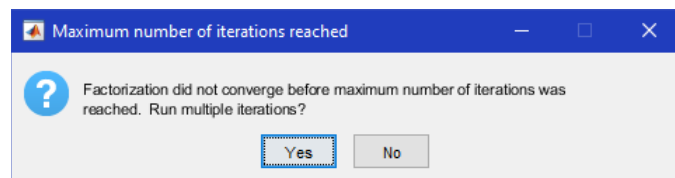
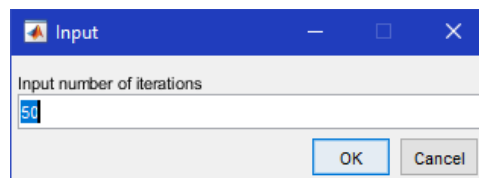
A screenshot of a dialog box titled "Maximum number of iterations reached". It features a question mark icon and the text: "Factorization did not converge before maximum number of iterations was reached. Run multiple iterations?". At the bottom are two buttons: "Yes" and "No".A screenshot of a dialog box titled "Input". It contains a text input field labeled "Input number of iterations" with the value "50" entered. At the bottom are two buttons: "OK" and "Cancel".

Figure 23. If one of the stopping criterion are not reached before the software reaches the specified number of iterations (Figure 16), the user will be asked if they want to run multiple iterations (left). If they select “Yes”, they will then be asked how many iterations they want to run (right). Once the specified number of iterations is reached, DZnmf2D will retain the run with the lowest Final Residual.

12. Saving figures and data

All plots may be exported by pressing the **Export Plots** pushbutton. These figures pop up in new windows and may be saved as vector images to be modified in a drafting program such as Adobe Illustrator. To save a figure as a vector image select File → Save as. Choose either .eps or .emf format

To export the NMF results press the **Export NMF Results** pushbutton. This will prompt a browser window for the user to name the file and select its destination (Figure 24). Results will be saved as a .xls file (Figure 24). If the optimization routine was used, all results from 2 to the maximum number of specified sources will be saved in individual sheets.

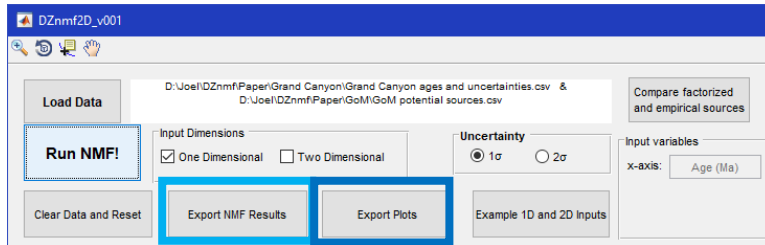


Figure 24. Figures can be exported for saving using the Export Plots function (dark blue box). NMF results can be saved to spreadsheets using the Export NMF Results (light blue box).

13. Comparing factorized and empirical sources

DZnmf2D includes the capability to compare factorized sources to a range of empirical sources in order to determine the empirical sources that most closely match the factorized sources. To use this function, select Compare factorized and empirical sources (Figure 25) after factorization is complete. In the browser window select a file with the potential empirical sources. This file should be formatted as the input files for factorization (section 4). Once the file is selected, the software will create distributions from the empirical sources using the same parameters that were used to create the distributions for factorization and compare them to the factorized distributions. It will output the three figures and two tables shown in Figures 26–28 for univariate distributions or Figures 29–31 for bivariate distributions.

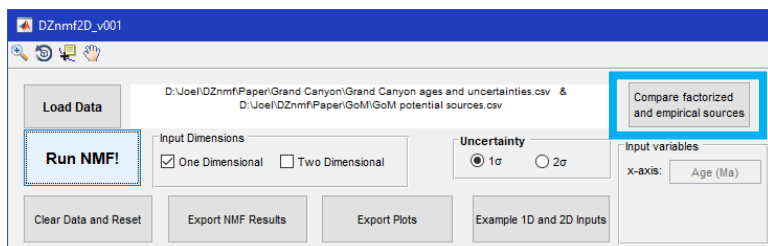


Figure 25. Figures can be exported for saving using the Export Plots function (dark blue box). NMF results can be saved to spreadsheets using the Export NMF Results (light blue box).

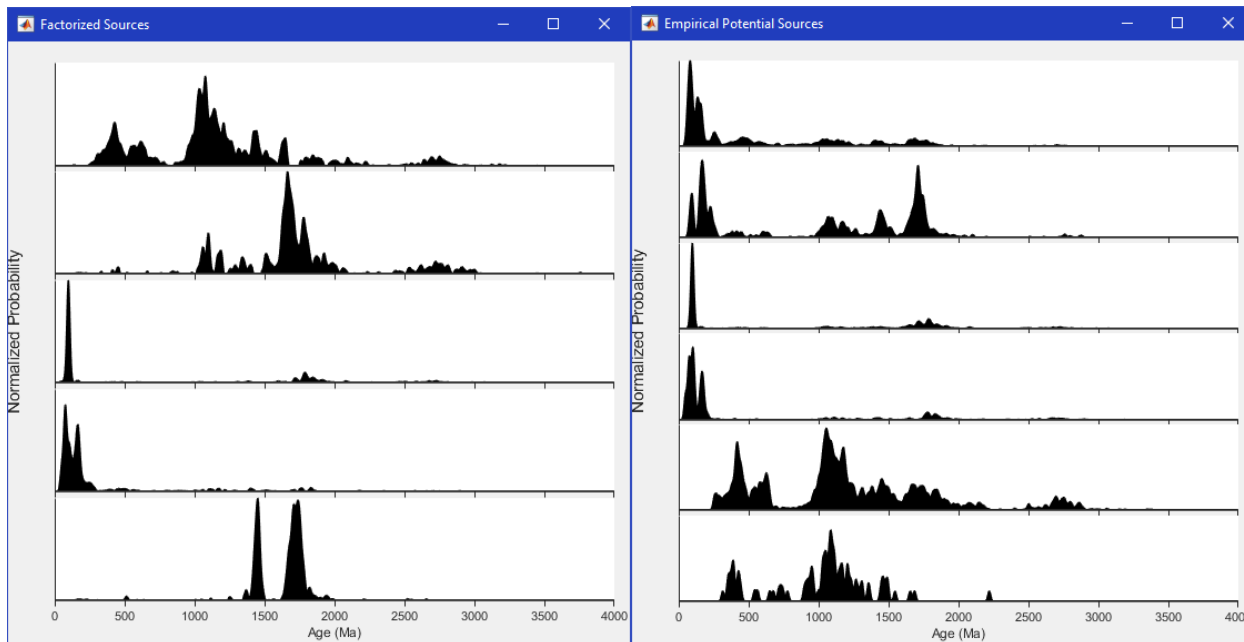


Figure 26. Factorized distributions (left) and the empirical distributions to which they will be compared (right).

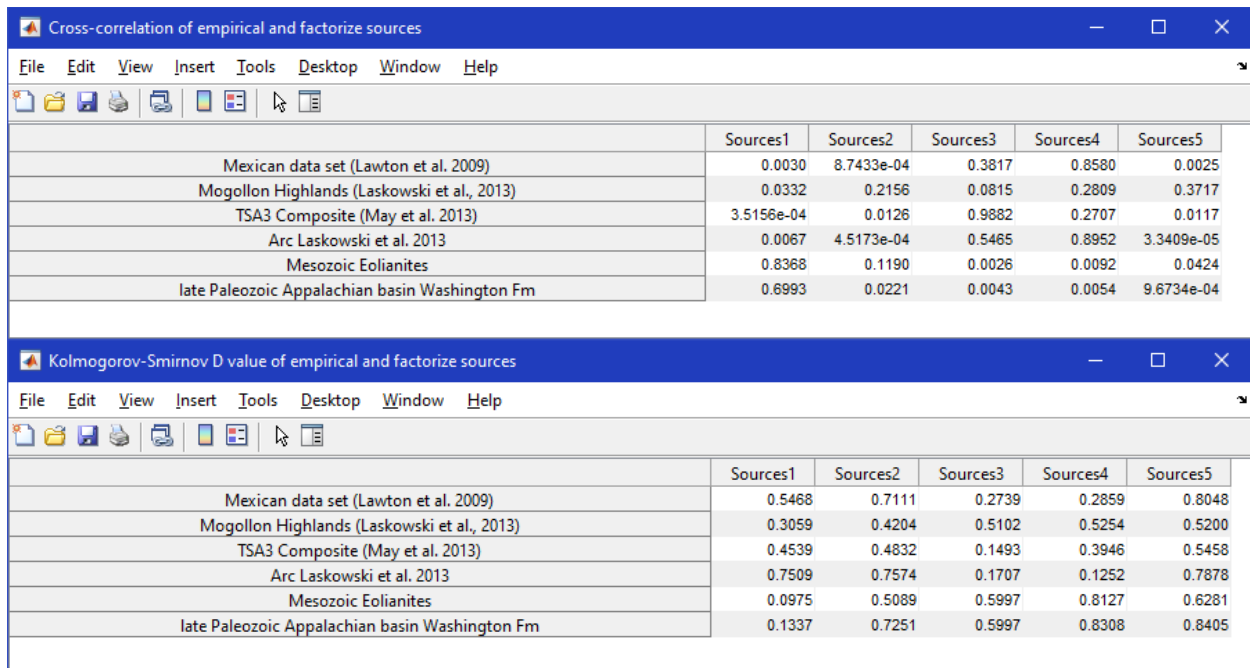


Figure 27. Comparison between factorized distributions (columns) and the empirical distributions to which they were compared (rows) using the Cross-correlation coefficient (top) and Kolmogorov-Smirnov D value (bottom).

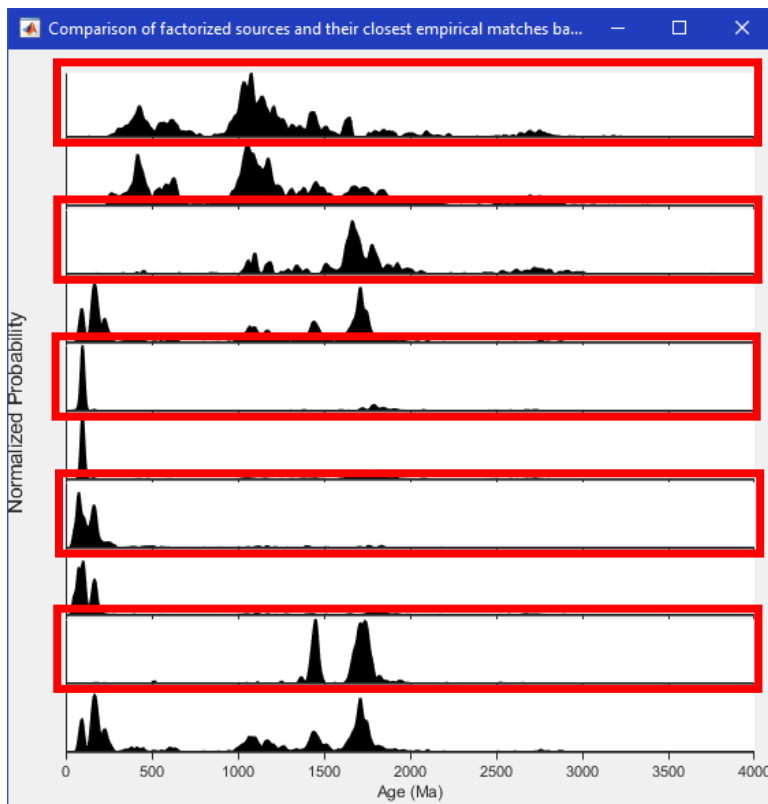


Figure 28. Stacked comparison of factorized distributions and their closest empirical matches based on Cross-correlation. Factorized sources are show as the first, third, fifth, seventh, and ninth from the top (red boxes), and empirical distributions are interspersed between them.

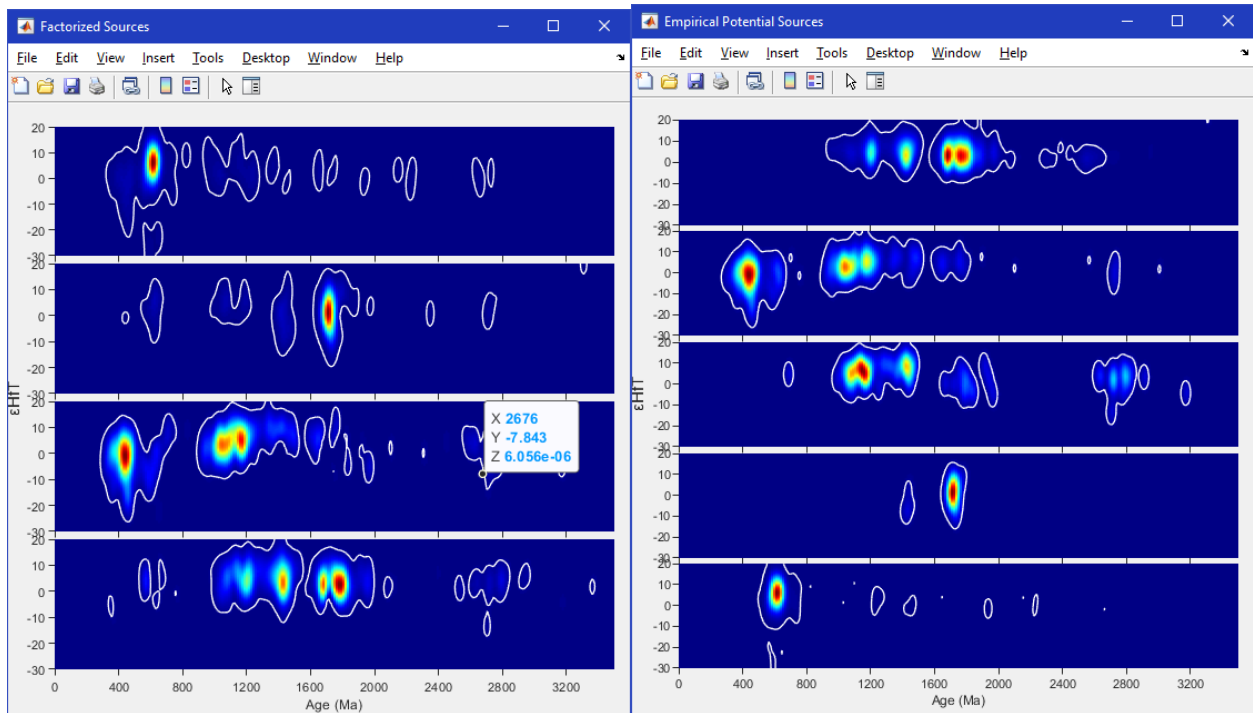


Figure 29. Factorized distributions (left) and the empirical distributions to which they will be compared (right).

Cross-correlation of empirical and factorize sources				
	Sources1	Sources2	Sources3	Sources4
Mojave (Wooden in Moj-Yav-Maz tab)	2.6534e-05	0.3749	0.0151	0.8905
Appalachian foreland (Thomas et al 2017 & SE US-App-Mueller tab)	0.0209	0.0022	0.9295	0.0579
Caddy and Mutual	4.9080e-05	0.0081	0.3725	0.3226
Upper Belt Supergroup (Missoula and Lemhi)	5.9780e-05	0.9842	4.3229e-04	0.1540
Avalonia	0.9892	0.0010	2.9002e-05	4.6914e-05

Kolmogorov-Smirnov D value of empirical and factorize sources				
	Sources1	Sources2	Sources3	Sources4
Mojave (Wooden in Moj-Yav-Maz tab)	0.8692	0.2852	0.6829	0.1648
Appalachian foreland (Thomas et al 2017 & SE US-App-Mueller tab)	0.5049	0.7823	0.0986	0.6338
Caddy and Mutual	0.8510	0.5350	0.4491	0.2596
Upper Belt Supergroup (Missoula and Lemhi)	0.9392	0.1178	0.8391	0.4816
Avalonia	0.1258	0.9025	0.6740	0.9249

Figure 30. Comparison between factorized distributions (columns) and the empirical distributions to which they were compared (rows) using the Cross-correlation coefficient (top) and Kolmogorov-Smirnov D value (bottom).

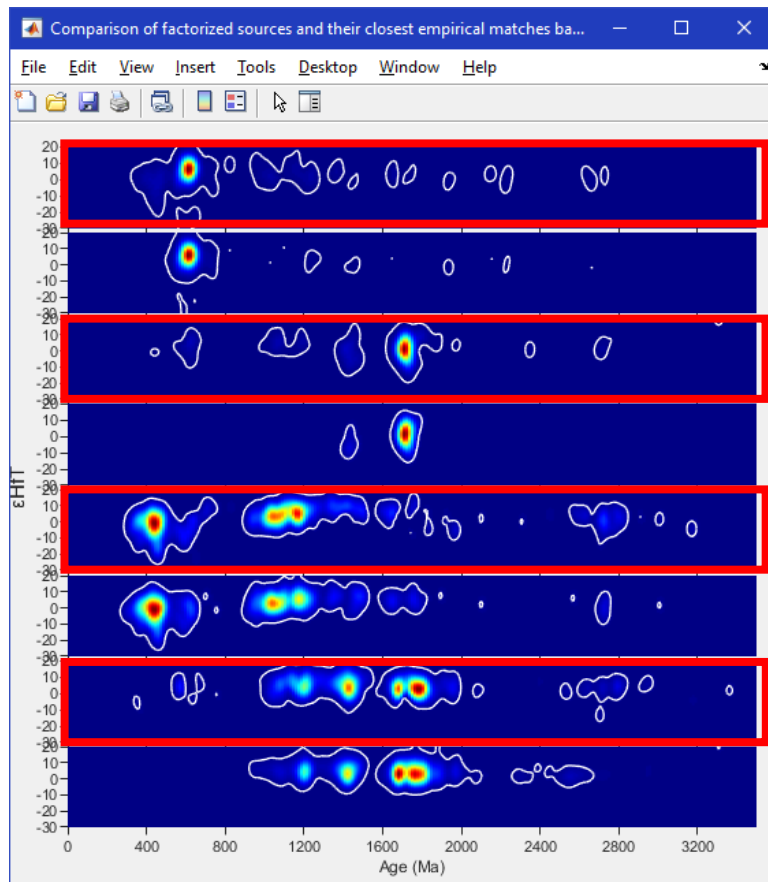


Figure 31. Stacked comparison of factorized distributions and their closest empirical matches based on Cross-correlation. Factorized sources are shown as the first, third, fifth, and seventh from the top (red boxes), and empirical distributions are interspersed between them.

Works Cited

Botev, Z.I., Grotowski, J.F., Kroese, D.P., 2010. Kernel density estimation via diffusion. *Annals of Statistics* 38, 2916–2957.

Saylor, J.E., Sundell, K.E., in review. Refined sediment source characterization via bivariate non-negative matrix factorization of detrital provenance indicators.

Saylor, J.E., Sundell, K.E., Sharman, G.R., 2019. Characterizing sediment sources by non-negative matrix factorization of detrital geochronological data. *Earth Planet. Sci. Lett.* 512, 46-58.