7. APPLICATION 5: DZNMF

Learning goals:

-How to determine the optimum number of sources for factorization using breakpoint analysis.

-How to factorize a sink sample set using DZnmf

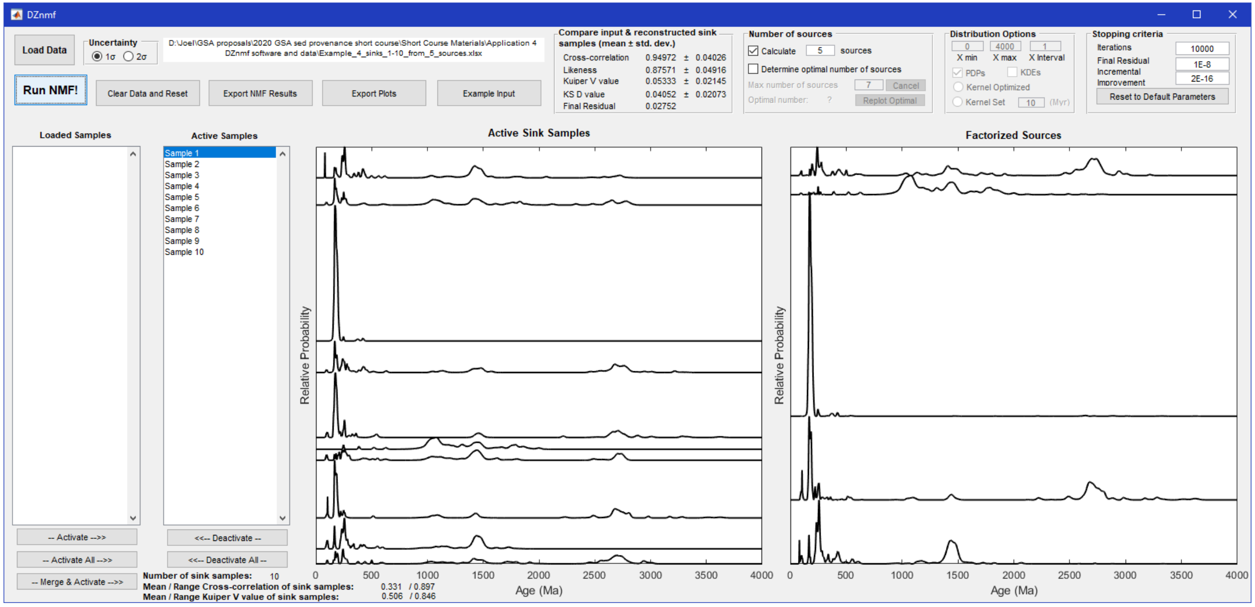
-Recognize the effect of the number of sinks on the quality of the source factorization

-Recognize the non-uniqueness of factorization and understand methods to mitigate the likelihood of identifying a local (rather than global) solution

7.1. Factorizing a synthetic data set

We will be using a couple of data sets that are mixtures of five source distributions.

1. Open DZnmf by double clicking the icon in in the folder where you saved the application.
2. Select "Load Data" and navigate to the file, "Application_5_sinks_1-10_from_5_sources.xlsx". Leave the options at their defaults.
3. Select "Activate All" at the bottom left.
4. Select "Run NMF!" After it finishes your window should look as below.



Uncertainty: ● 1σ ○ 2σ

D:\Joel\GSA proposals\2020 GSA sed provenance short course\Short Course Materials\Application 4 DZnmf software and data\Example_4_sinks_1-10_from_5_sources.xlsx

Compare input & reconstructed sink samples (mean ± std. dev.)

Cross-correlation	0.94972	±	0.04026
Likeness	0.87571	±	0.04916
Kuiper V value	0.05333	±	0.02145
KS D value	0.04052	±	0.02073
Final Residual	0.02752		

Number of sources

☑ Calculate 5 sources

☐ Determine optimal number of sources

Max number of sources 7

Optimal number: ?

Distribution Options

0 X min, 4000 X max, 1 X Interval

☑ PDPs ☐ KDEs

○ Kernel Optimized

○ Kernel Set 10 (Myr)

Stopping criteria

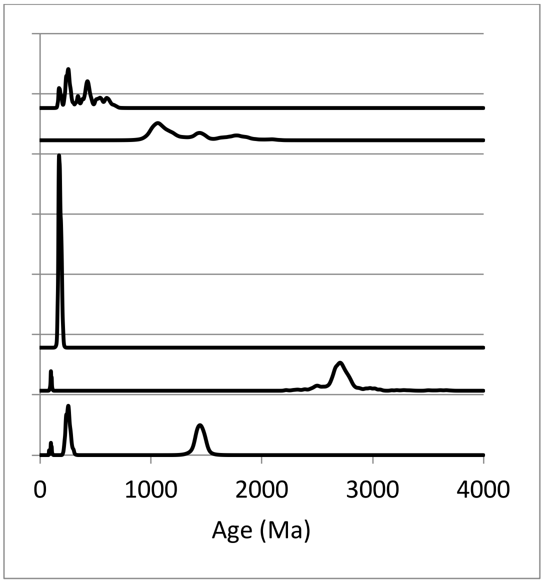
Iterations	10000
Final Residual	1E-8
Incremental Improvement	2E-16

Number of sink samples: 10

Mean / Range Cross-correlation of sink samples: 0.331 / 0.897

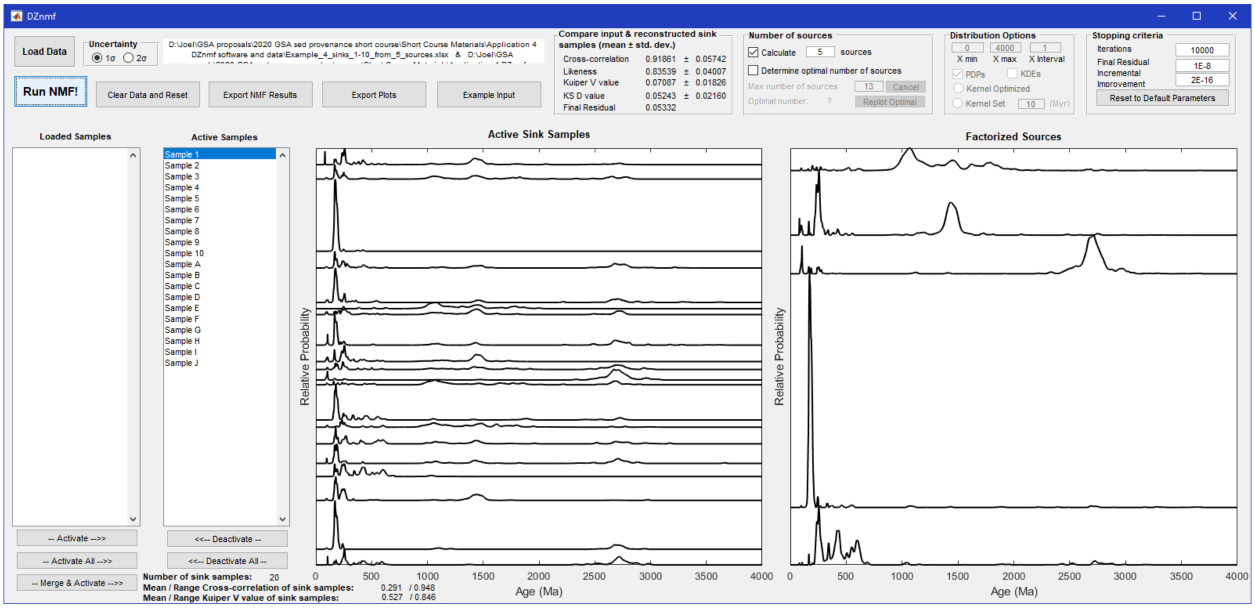
Mean / Range Kuiper V value of sink samples: 0.506 / 0.846

Because we created these samples we know what the original sources looked like (right). Clearly the algorithm is doing a poor job of identifying these sources. The topmost and second from bottom sources are particularly poorly separated.



7.2. Impact of the number of samples on factorization

1. Let’s increase the number of samples that we are including in the analysis. Without clearing the data from the previous step, select “Load Data” and navigate to the file, “Application_5_sinks_A-J_from_5_sources.xlsx”. Leave the options at their defaults.
2. Select “Activate All” at the bottom left.
3. Note that we now have 20 active samples (samples 1–10 and A–J). DZnmf can merge data sets on the fly to maximize the ability to explore data sets. You can also select multiple samples in the “Loaded Samples” list and merge them into one sample in the “Active Samples” list by highlighting them, and selecting “Merge & Activate” at the bottom left.
4. Select “Run NMF!” After it finishes your window should look as below.



Uncertainty ● 1σ ○ 2σ

D:\Joel\GSA proposals\2020 GSA sed provenance short course\Short Course Materials\Application 4 DZnmf software and data\Example_4_sinks_1-10_from_5_sources.xlsx & D:\Joel\GSA

Compare input & reconstructed sink samples (mean ± std. dev.)

Cross-correlation	0.91861	±	0.05742
Likeness	0.83539	±	0.04007
Kuiper V value	0.07087	±	0.01826
KS D value	0.05243	±	0.02160
Final Residual	0.05332		

Number of sources
☑ Calculate 5 sources
□ Determine optimal number of sources
Max number of sources 13
Optimal number: ?

Distribution Options
0 X min 4000 X max 1 X Interval
☑ PDPs □ KDEs
○ Kernel Optimized
○ Kernel Set 10 (Myr)

Stopping criteria
Iterations 10000
Final Residual 1E-8
Incremental Improvement 2E-16

Active Samples: Sample 1, Sample 2, Sample 3, Sample 4, Sample 5, Sample 6, Sample 7, Sample 8, Sample 9, Sample 10, Sample A, Sample B, Sample C, Sample D, Sample E, Sample F, Sample G, Sample H, Sample I, Sample J

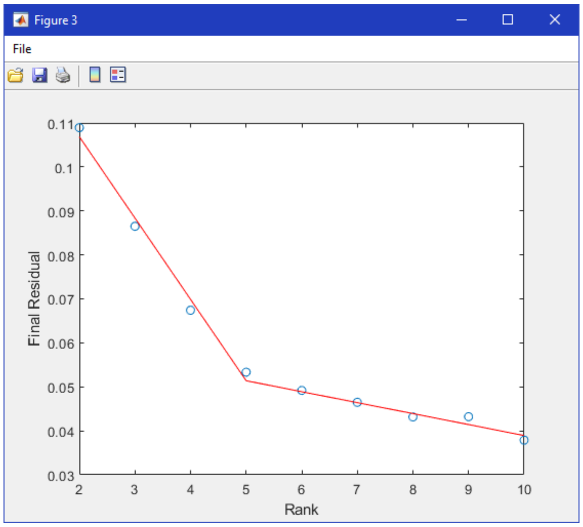
Number of sink samples: 20
Mean / Range Cross-correlation of sink samples: 0.291 / 0.948
Mean / Range Kuiper V value of sink samples: 0.527 / 0.846

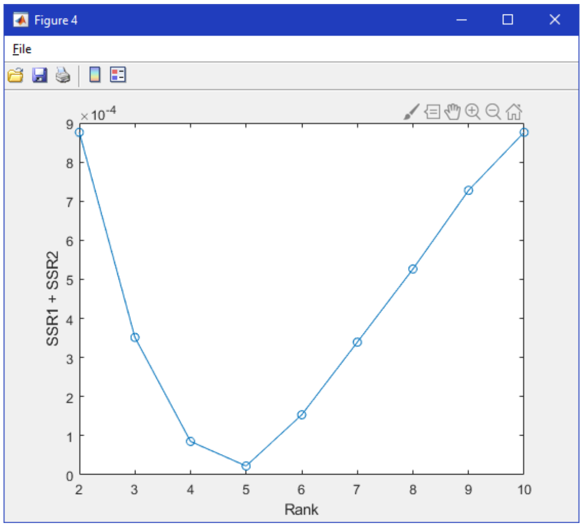
5. Comparing the new results to the known age distributions above shows two things. First, the **order** of the results does not matter. Secondly, and more importantly, adding more samples resulted in much more discriminatory power and a closer match between the known source distributions and the factorized distributions.

7.3. Determining the optimum number of sources

This would usually be the first step in an NMF analysis.

1. Without clearing the data, select “Determine optimal number of sources”.
2. Set the “Max number of sources” to 10.
3. Select “Run NMF!” again. The algorithm will cycle through a factorization of all ranks from 2–10 and calculate the final residual. It should take about 10 minutes.
4. When it finishes, two windows will open, as shown below. The plot on the left shows the Final Residual for each rank between 2 and 10 as blue data points and the optimized segmented linear regression as a red line. The plot on the right shows the sum of summed squared residuals (SSSR) based on the segmented linear regression. The optimum rank is the one that minimizes the SSSR. Note that the calculated optimum rank (5) coincides with the known number of sources.

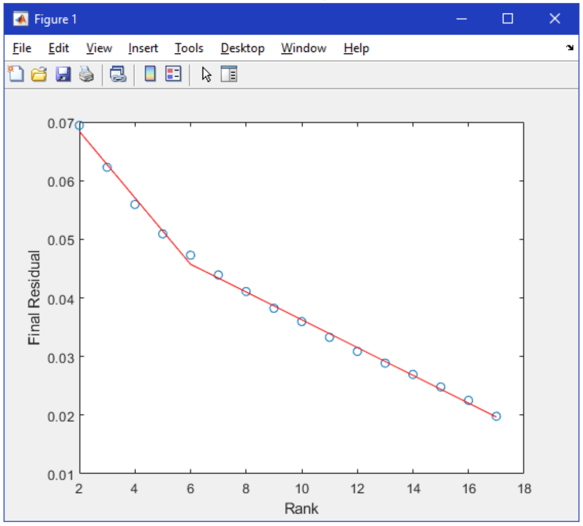


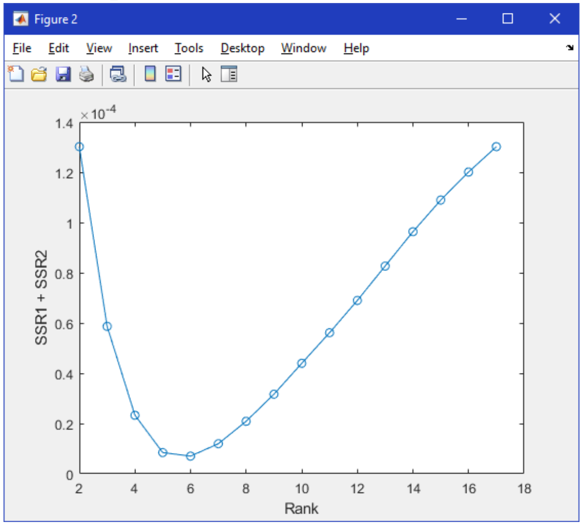


7.4. Non-negative matrix factorization of an empirical data set

We will be using a couple of empirical data sets from Neoproterozoic–Triassic strata from western Laurentia (Gehrels et al., 2011; Gehrels and Pecha, 2014) to explore some of the complexities of using DZnmf.

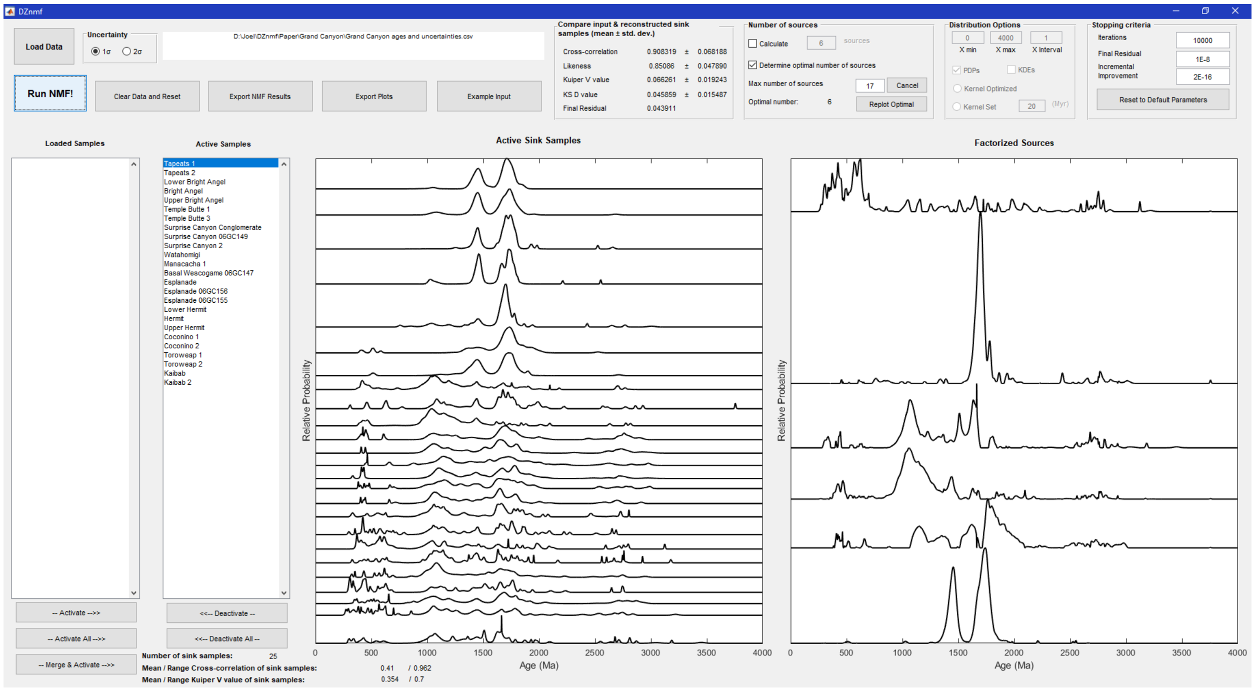
1. Select “Load Data” and navigate to the file, “Application_5_Gehrels_et_al_2011_Grand_Canyon_U-Pb.xlsx”. Leave the options at their defaults. Uncertainty is reported as “1sigma”, and we will visualize the samples as PDPs.
2. Select “Activate All” at the bottom left.
3. Usually the first step would be to factorize the data set to a wide range of ranks to determine the optimum rank. Because this process takes about 20 minutes, I have done it already and show the results below. The optimum rank is 6.





4. Under “Number of sources” select the checkbox for “Calculate” and input “6” in the sources box.
5. Select “Run NMF!”

6. Once the factorization is complete, your DZnmf window should be as below. Note that the order of the sources may not match this figure. Compare these factorized sources to the empirical sources reported by Saylor et al. (2019).



Uncertainty: ● 1σ ○ 2σ

D:\Joel\DZnmf\Paper\Grand Canyon\Grand Canyon ages and uncertainties.csv

Compare input & reconstructed sink samples (mean ± std. dev.)

Cross-correlation	0.908319	±	0.068188
Likeness	0.85086	±	0.047890
Kuiper V value	0.066261	±	0.019243
KS D value	0.045859	±	0.015487
Final Residual	0.043911		

Number of sources
☐ Calculate 6 sources
☑ Determine optimal number of sources
Max number of sources 17
Optimal number: 6

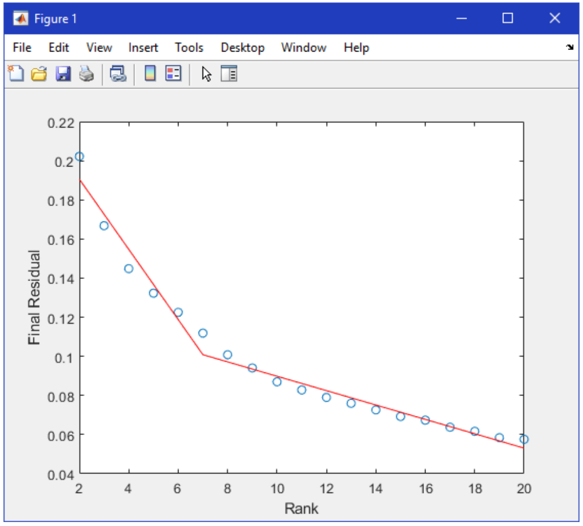
Distribution Options
0 X min
4000 X max
1 X Interval
☑ PDPs ☐ KDEs
○ Kernel Optimized
○ Kernel Set 20 (Myr)

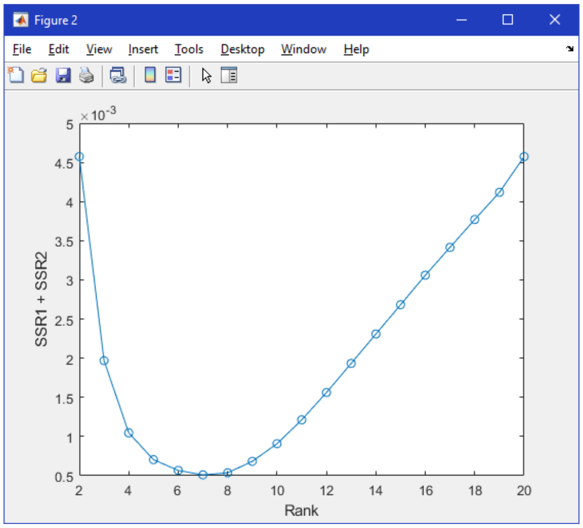
Stopping criteria
Iterations 10000
Final Residual 1E-8
Incremental Improvement 2E-16

Active Samples:
Tapeats 1
Tapeats 2
Lower Bright Angel
Bright Angel
Upper Bright Angel
Temple Butte 1
Temple Butte 3
Surprise Canyon Conglomerate
Surprise Canyon 06GC149
Surprise Canyon 2
Watahomigi
Manacacha 1
Basal Wescogame 06GC147
Esplanade
Esplanade 06GC156
Esplanade 06GC155
Lower Hermit
Hermit
Upper Hermit
Coconino 1
Coconino 2
Toroweap 1
Toroweap 2
Kaibab
Kaibab 2

Number of sink samples: 25
Mean / Range Cross-correlation of sink samples: 0.41 / 0.962
Mean / Range Kuiper V value of sink samples: 0.354 / 0.7

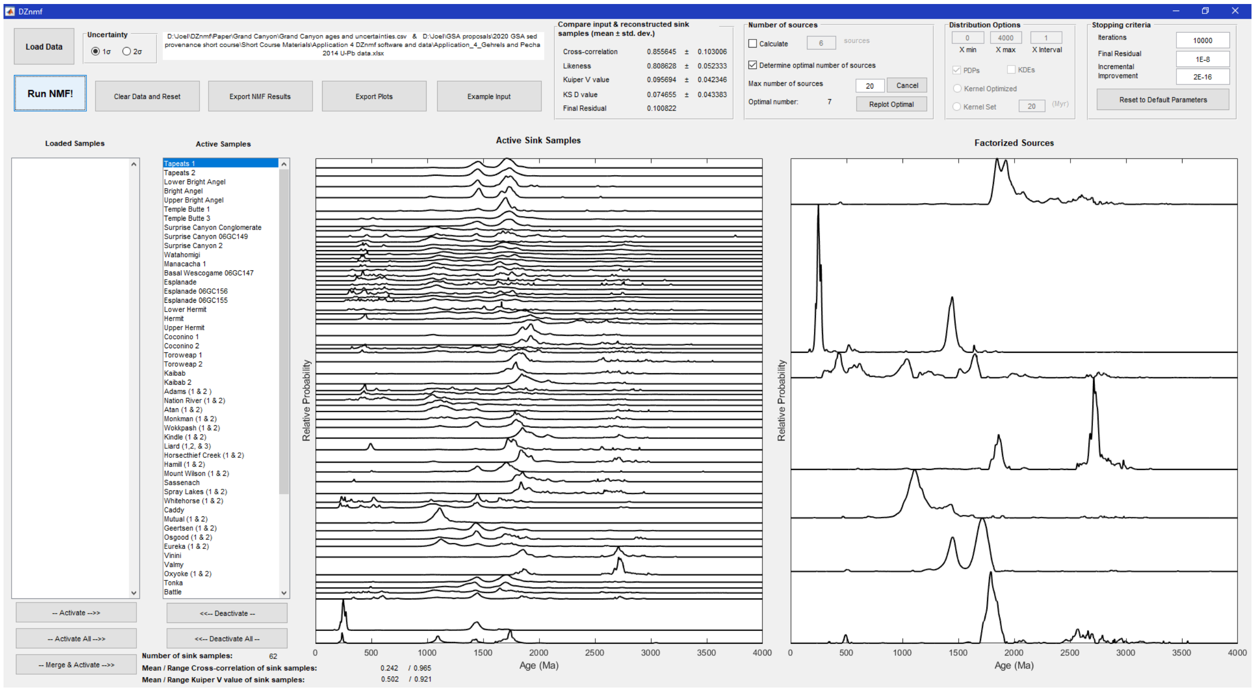
7. We will take one more step: adding in additional data and repeating the factorization.
8. Select "Deactivate All"
9. Select "Load Data" and navigate to the file, "Application_5_Gehrels_and_Pecha_2014_W_Laurentia_U-Pb.xlsx". Leave the options at their defaults. Uncertainty is reported as "1sigma", and we will visualize the samples as PDPs.





10. As above, we would usually factorize the data set to a wide range of ranks to determine the optimum rank. I have done it already and show the results above. The optimum rank is 7.

11. Under “Number of sources” select the checkbox for “Calculate” and input “7” in the sources box.
12. Select “Run NMF!”
13. Once the factorization is complete, your DZnmf window should be as below. Note that the order of the sources may not match this figure. Compare these results to the results from step 6. From bottom (1) to top (7):
    a. Step 13 Source 1 => Step 6 Source 5
    b. Step 13 Source 2 => Step 6 Source 1
    c. Step 13 Source 3 => Step 6 Source 3
    d. Step 13 Source 4 => Novel
    e. Step 13 Source 5 => Step 6 Sources 4 & 6
    f. Step 13 Source 6 => Novel
    g. Step 13 Source 7 => Step 6 Source 2?



Uncertainty ● 1σ ○ 2σ

D:\Joel\DZnmf\Paper\Grand Canyon\Grand Canyon ages and uncertainties.csv & D:\Joel\GSA proposals\2020 GSA sed provenance short course\Short Course Materials\Application 4 DZnmf software and data\Application_4_Gehrels and Pecha 2014 U-Pb data.xlsx

Compare input & reconstructed sink samples (mean ± std. dev.)

Cross-correlation	0.855645	±	0.103006
Likeness	0.808628	±	0.052333
Kuiper V value	0.095694	±	0.042346
KS D value	0.074655	±	0.043383
Final Residual	0.100822		

Number of sources
□ Calculate 6 sources
☑ Determine optimal number of sources
Max number of sources 20
Optimal number: 7

Distribution Options
0 X min
4000 X max
1 X Interval
☑ PDPs □ KDEs
○ Kernel Optimized
○ Kernel Set 20 (Myr)

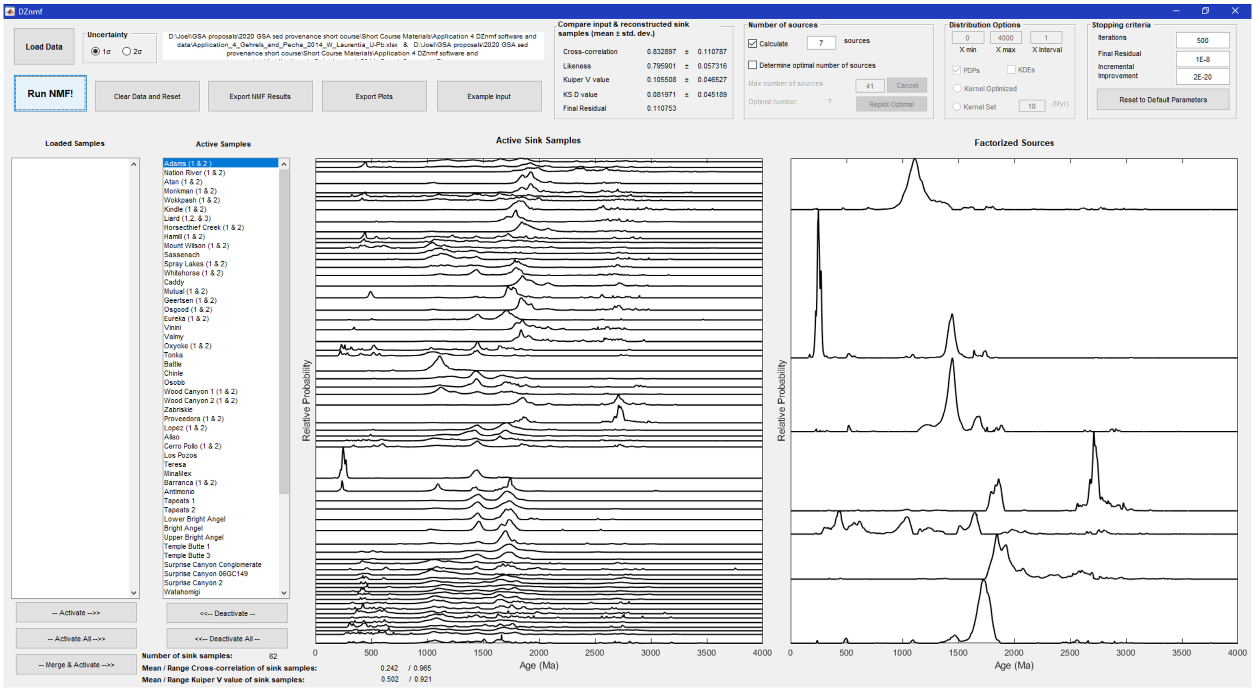
Stopping criteria
Iterations 10000
Final Residual 1E-8
Incremental Improvement 2E-16

Active Samples:
Tapeats 1
Tapeats 2
Lower Bright Angel
Bright Angel
Upper Bright Angel
Temple Butte 1
Temple Butte 3
Surprise Canyon Conglomerate
Surprise Canyon 06GC149
Surprise Canyon 2
Watahomigi
Manacacha 1
Basal Wescogame 06GC147
Esplanade
Esplanade 06GC156
Esplanade 06GC155
Lower Hermit
Hermit
Upper Hermit
Coconino 1
Coconino 2
Toroweap 1
Toroweap 2
Kaibab
Kaibab 2
Adams (1 & 2)
Nation River (1 & 2)
Atan (1 & 2)
Monkman (1 & 2)
Wokkpash (1 & 2)
Kindle (1 & 2)
Liard (1,2, & 3)
Horsecthief Creek (1 & 2)
Hamill (1 & 2)
Mount Wilson (1 & 2)
Sassenach
Spray Lakes (1 & 2)
Whitehorse (1 & 2)
Caddy
Mutual (1 & 2)
Geertsen (1 & 2)
Osgood (1 & 2)
Eureka (1 & 2)
Vinini
Valmy
Oxyoke (1 & 2)
Tonka
Battle

Number of sink samples: 62
Mean / Range Cross-correlation of sink samples: 0.242 / 0.965
Mean / Range Kuiper V value of sink samples: 0.502 / 0.921

14. One final point. I ran the same data set through a factorization with only 500 iterations and got the result below. Notice that the age distributions are different from above. Notice also that the factorization below has a higher final residual and lower mean cross-correlation coefficient. NMF is non-unique. We have tried to account for this by including the possibility of running multiple trials to maximize the probability of finding a global minimum in final residual.
15. In order to use this feature, set both the final residual criteria and incremental improvement criteria to extremely low values (i.e., 1E-20). The algorithm will factorize the data set but if it reaches the target iterations before it reaches the ending

criteria, it will ask if you want to repeat the factorization (see below). After running the prescribed number of factorizations it will select the one with the lowest final residual. We recommend repeating the factorization at least 50 times prior to accepting the results.



Uncertainty ● 1σ ○ 2σ

D:\Joel\GSA proposals\2020 GSA sed provenance short course\Short Course Materials\Application 4 DZnmf software and data\Application_4_Gehrels_and_Pecha_2014_W_Laurentia_U-Pb.xlsx & D:\Joel\GSA proposals\2020 GSA sed provenance short course\Short Course Materials\Application 4 DZnmf software and

Compare input & reconstructed sink samples (mean ± std. dev.)

```
Cross-correlation   0.832897  ±  0.110787
Likeness            0.795901  ±  0.057316
Kuiper V value      0.105508  ±  0.046527
KS D value          0.081971  ±  0.045189
Final Residual      0.110753
```

Number of sources

☑ Calculate 7 sources

☐ Determine optimal number of sources

Max number of sources 41

Optimal number: ?

Distribution Options

X min 0 X max 4000 X Interval 1

☑ PDPs ☐ KDEs

○ Kernel Optimized

○ Kernel Set 10 (Myr)

Stopping criteria

Iterations 500

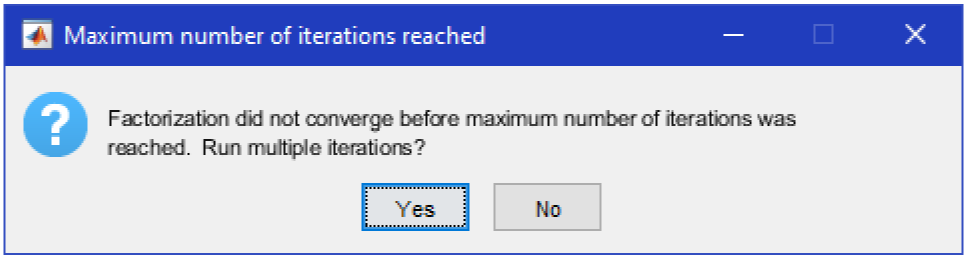
Final Residual 1E-8

Incremental Improvement 2E-20

Active Samples:

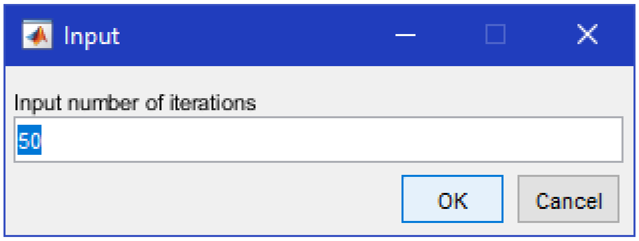
Adams (1 & 2)
Nation River (1 & 2)
Atan (1 & 2)
Monkman (1 & 2)
Wokkpash (1 & 2)
Kindle (1 & 2)
Liard (1,2, & 3)
Horsecthief Creek (1 & 2)
Hamill (1 & 2)
Mount Wilson (1 & 2)
Sassenach
Spray Lakes (1 & 2)
Whitehorse (1 & 2)
Caddy
Mutual (1 & 2)
Geertsen (1 & 2)
Osgood (1 & 2)
Eureka (1 & 2)
Vinini
Valmy
Oxyoke (1 & 2)
Tonka
Battle
Chinle
Osobb
Wood Canyon 1 (1 & 2)
Wood Canyon 2 (1 & 2)
Zabriskie
Proveedora (1 & 2)
Lopez (1 & 2)
Aliso
Cerro Pollo (1 & 2)
Los Pozos
Teresa
MinaMex
Barranca (1 & 2)
Antimonio
Tapeats 1
Tapeats 2
Lower Bright Angel
Bright Angel
Upper Bright Angel
Temple Butte 1
Temple Butte 3
Surprise Canyon Conglomerate
Surprise Canyon 06GC149
Surprise Canyon 2
Watahomigi

```
Number of sink samples:                          62
Mean / Range Cross-correlation of sink samples:  0.242  / 0.965
Mean / Range Kuiper V value of sink samples:     0.502  / 0.921
```



Maximum number of iterations reached

Factorization did not converge before maximum number of iterations was reached. Run multiple iterations?



Input

Input number of iterations

50

Works Cited

Gehrels, G., Pecha, M., 2014. Detrital zircon U-Pb geochronology and Hf isotope geochemistry of Paleozoic and Triassic passive margin strata of western North America. Geosphere 10, 49–65.

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, 183–200.

Kuiper, N.H., 1960. Tests concerning random points on a circle. Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen 63, 38–47.

Licht, A., Pullen, A., Kapp, P., Abell, J., Giesler, N., 2016. Eolian cannibalism: Reworked loess and fluvial sediment as the main sources of the Chinese Loess Plateau. Geol. Soc. Am. Bull. 128, 944–956.

Massey Jr, F.J., 1951. The Kolmogorov-Smirnov test for goodness of fit. Journal of the American statistical Association 46, 68-78.

Saylor, J.E., Knowles, J.N., Horton, B.K., Nie, J.S., Mora, A., 2013. Mixing of Source Populations Recorded in Detrital Zircon U-Pb Age Spectra of Modern River Sands. J. Geol. 121, 17–33.

Saylor, J.E., Stockli, D.F., Horton, B.K., Nie, J., Mora, A., 2012. Discriminating rapid exhumation from syndepositional volcanism using detrital zircon double dating: Implications for the tectonic history of the Eastern Cordillera, Colombia. Geol. Soc. Am. Bull. 124, 762–779.

Saylor, J.E., Sundell, K.E., 2016. Quantifying comparison of large detrital geochronology data sets. Geosphere 12, 203–220.

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.

Sharman, G.R., Sharman, J.P., Sylvester, Z., 2018. detritalPy: A Python-based toolset for visualizing and analysing detrital geo-thermochronologic data. The Depositional Record 4, 202-215.

Sundell, K.E., Saylor, J.E., 2017. Unmixing detrital geochronology age distributions. Geophysics, Geochemistry, Geosystems 18.