Matrix-Tensor Factorization Preliminary Results

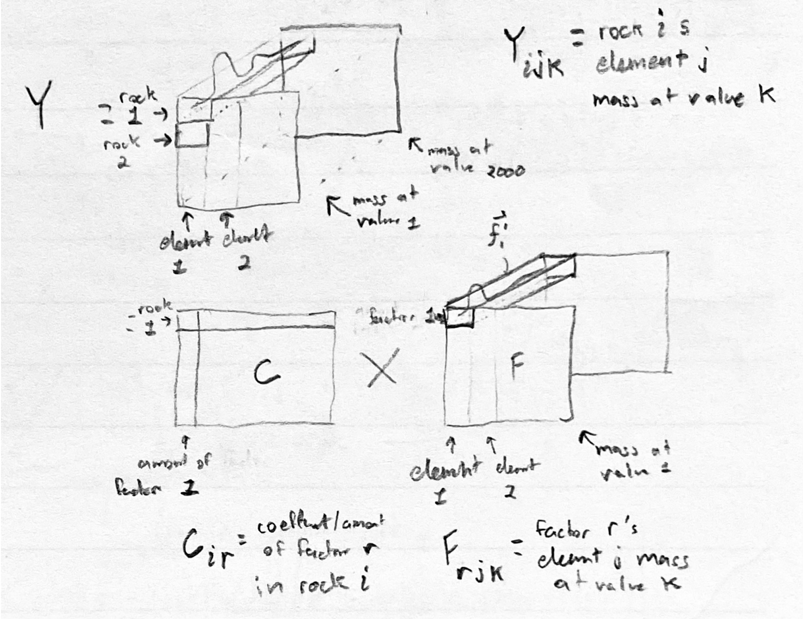
#sediment-analysis #research

Model

The 2D model is given below:

$$
\begin{array}{c}
\text{rock 1}\left\{\begin{array}{l}\text{grain 1}\rightarrow\\ \text{grain 2}\rightarrow\\ \vdots\end{array}\right. \\
\text{rock 2}\left\{\begin{array}{l}\text{grain 1}\rightarrow\\ \text{grain 2}\rightarrow\\ \vdots\end{array}\right.
\end{array}
\begin{array}{c}
\begin{bmatrix} g^1_{11} & g^1_{12} & \cdots \\ g^1_{21} & g^1_{22} & \cdots \\ \vdots & \vdots & \\ \hline g^2_{11} & g^2_{12} & \cdots \\ g^2_{21} & g^2_{22} & \cdots \\ \vdots & \vdots & \end{bmatrix} \\
\begin{array}{ccc}\uparrow & \uparrow & \\ \text{element 1} & \text{element 2} & \cdots\end{array}
\end{array}
=
\begin{array}{c}
\begin{bmatrix} a_1 & \\ & \ddots \end{bmatrix}\\
\begin{array}{cc}\uparrow & \\ \text{amount of grain 1} & \cdots\end{array}
\end{array}
\begin{array}{c}
\begin{bmatrix} c^1_{11} & c^1_{12}\\ c^1_{21} & c^1_{22}\\ \vdots & \vdots \\ \hline c^2_{11} & c^2_{12} \\ c^2_{21} & c^2_{22}\\ \vdots & \vdots\end{bmatrix}\\
\begin{array}{cc}\uparrow & \uparrow\\ \text{factor 1 coefficients} & \text{factor 2 coefficients}\end{array}
\end{array}
\begin{array}{c}
\begin{bmatrix} f_{11} & f_{12} & \cdots \\ f_{21} & f_{22} & \cdots\end{bmatrix}\begin{array}{l}\leftarrow \text{factor 1}\\ \leftarrow\text{factor 2}\end{array}\\
\begin{array}{ccc}\uparrow & \uparrow & \\ \text{element 1} & \text{element 2} & \cdots\end{array}
\end{array}
$$

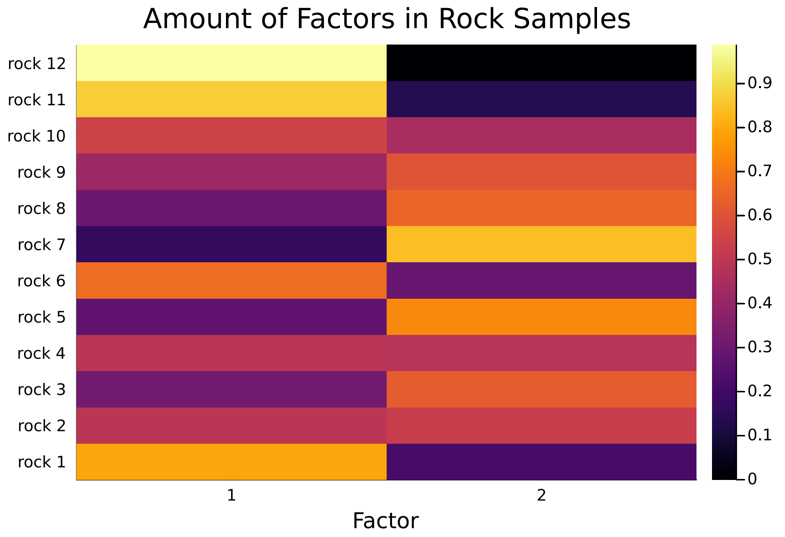
For the 3D model, we replace each grain measurement with a probability distribution extending a third dimension by colapsing the grains within a rock to get a distribution of values for that rock. The also extends the factor matrix to a factor tensor. Here is a rough sketch of the factorization $Y_{ijk} = \sum_{r=1}^{R} C_{ir} F_{jrk}$.

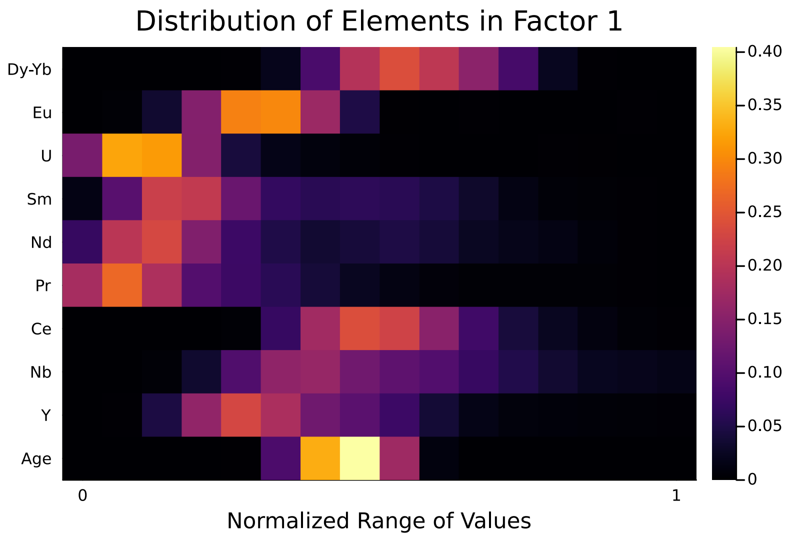


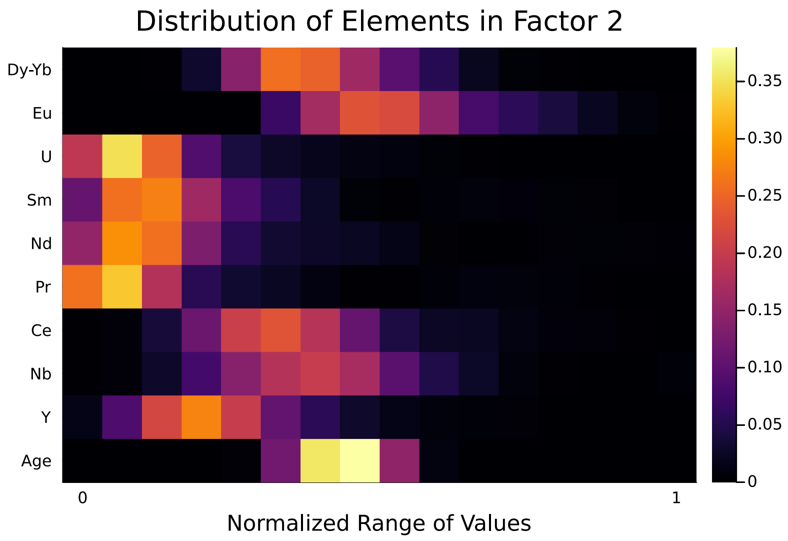
Results

Following the matrix-tensor factorization, we **do** obtain reasonable looking coefficients and factor distributions!

Below are the results using the Lee et. al. 2021 density function data. The first plot is the learned matrix C and the next two are the 2 horizontal slices of F.







You can look at these plots and notice a few things. In the first plot of the coefficient matrix C, you can see rock 12 has mostly factor 1 whereas rock 7 is mostly factor 2. Looking at the horizontal slices of the factor tensor F, we see factor 1 is more likely to have more Dy, Ce, and Y, whereas factor 2 is more likely to have more Eu. With the number of grain samples given, is looks like U, Sm, Nd, Pr, Nb, and Age distributions are similar across all the rocks.

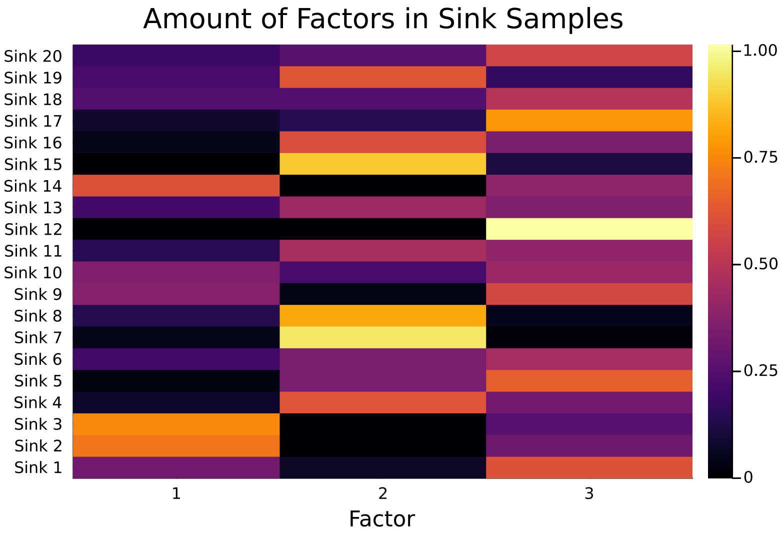
Evaluation

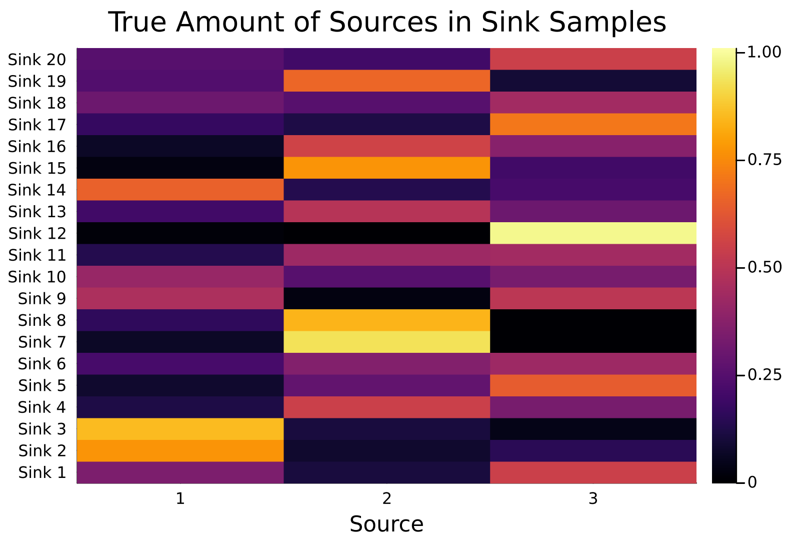
It is important to note that the factorization is not perfect, since the relative root-squared error is 14% $(\mathrm{RSE} = \left\|Y - \hat{Y}\right\|_F)$ but the relative root-mean-squared-error is 0.32% $(\mathrm{RMSE} = \frac{\left\|Y-\hat{Y}\right\|_F}{\sqrt{N}}$, N is the number of entries in Y). Similarly, the relative MAE is 14.6% $(\mathrm{MAE} = \frac{\left\|Y-\hat{Y}\right\|_1}{N}$). Note to make the quantity "relative", we divide by the norm of Y (Frobenius or 1-norm respectively).

Using 20 sinks from 3 sources data

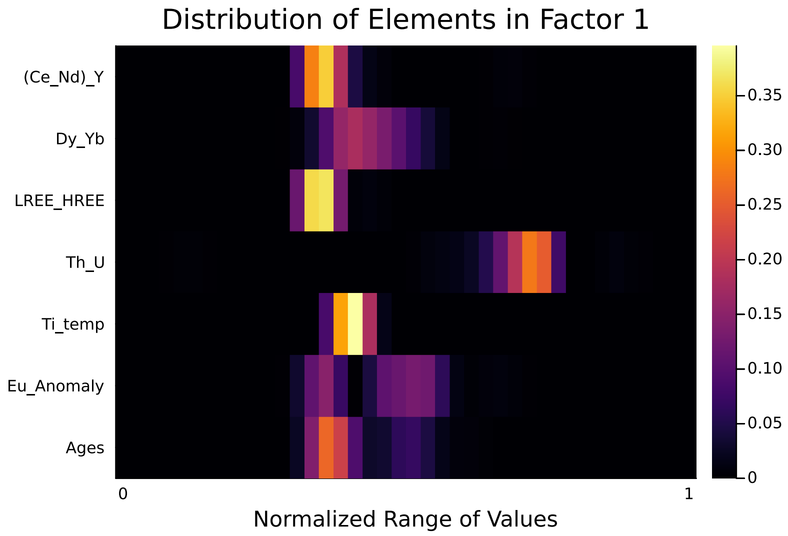
Below are the results using the Sundall et. al. 2022 density function data.

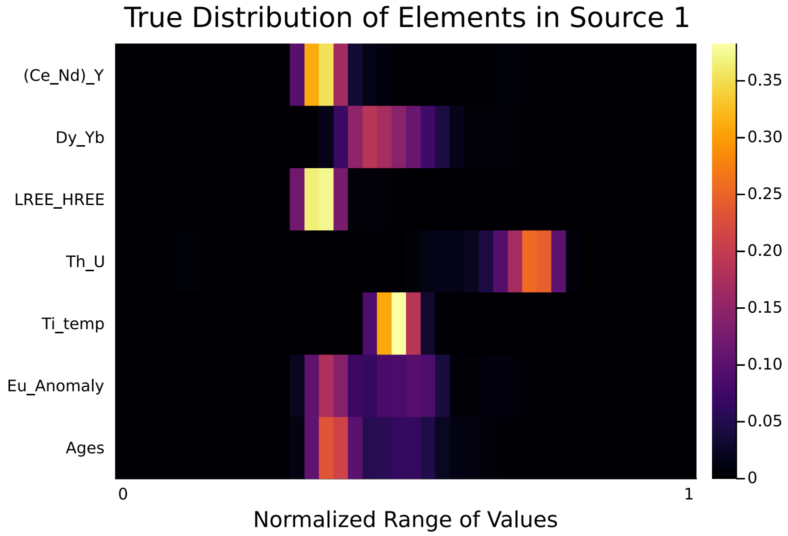
Note the 2nd and 3rd *true* sources were swapped to match the learned ordering of sources.

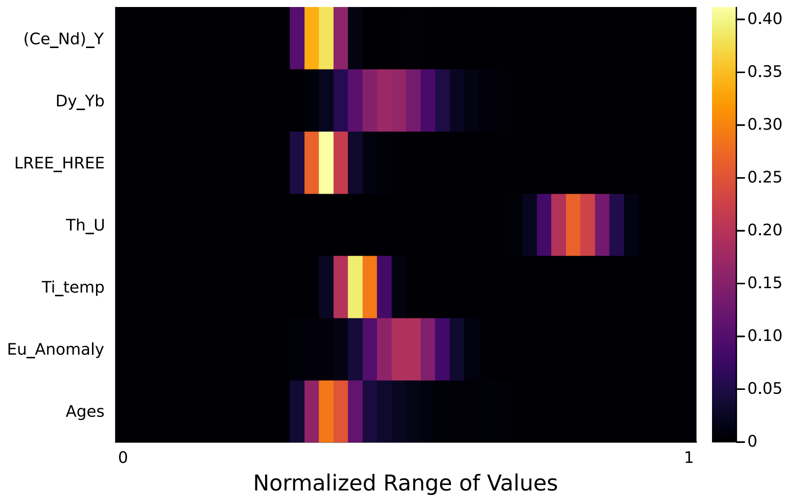


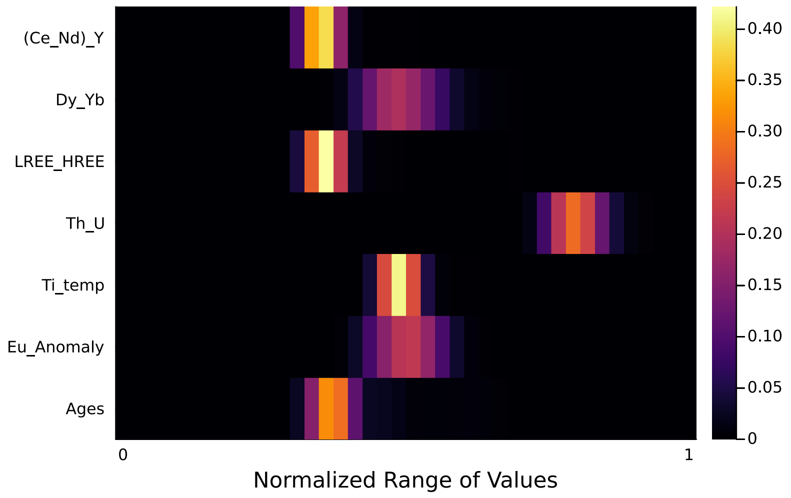


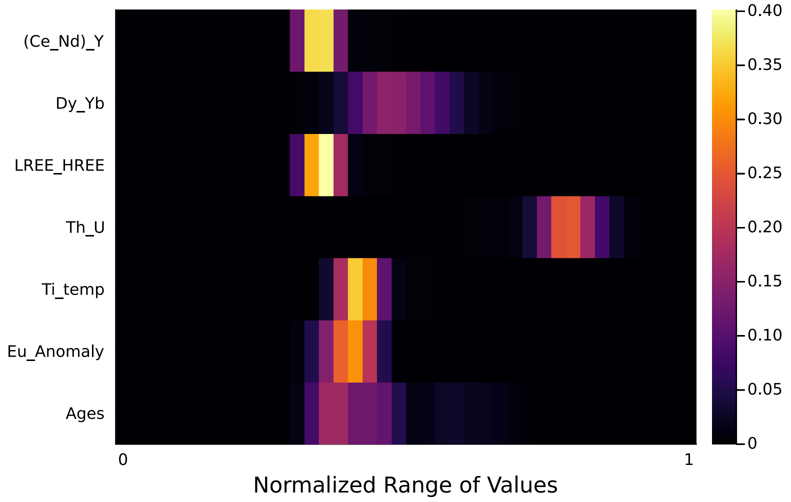
And the factors:

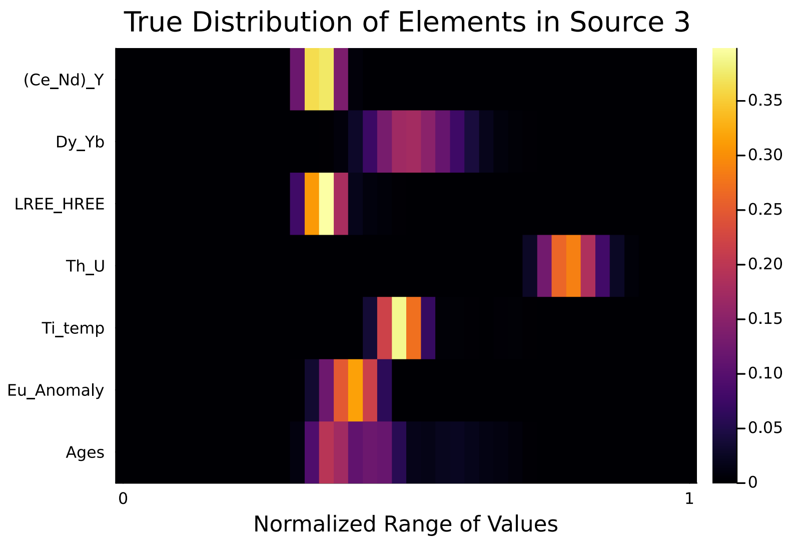





Distribution of Elements in Factor 2


True Distribution of Elements in Source 2


Distribution of Elements in Factor 3



From this, it looks like Th and Eu are the biggest characterizers. And the only miss was Ti where the distributions should be swapped between source 2 and 3. The factorization nonetheless has a relative RSE of only 6% and relative RMSE is 0.08%.

I imagine this would perform better if the full mixed distributions took up the whole x-range of data. Many of the distributions incorrectly spill into the negatives values because of the kernel estimation, and most do not use the range of data to the right. This means that when smoothing, some elements' distributions look identical between sinks.

To Do:

- ☑ ~~make ground truth plots of factors' distributions~~
- ☑ ~~normalize data so the row sums of distributions is 1 before factorization.~~
- ☐ squish range to use the full width of data, and remove negative values of ppm/ages where it does not make sense
- ☐ auto-swap the learned sources to match true sources for easier comparison
- ☐ Use https://github.com/JuliaStats/Distances.jl to evaluate closeness of results/error in computation
- ☐ Estimate the distributions using https://github.com/JuliaStats/KernelDensity.jl to improve detail of density functions
- ☐ Estimate which learned source each grain belongs to by labelling it with the highest likelihood estimate between the grain measurements and the learned distributions of each source

It should be noted that this method is very promising *because* the distribution estimation is not sophisticated. We'd expect a smarter method to decompose more accurately and display clearer results.