Course Schedule

- Introduction
- 1. Data visualization: PDPs, KDEs, and CDFs
- 2. detritalPy
    - Break
- 3. Statistical metrics & MDS
- 4. DZmds & Dzstats application
    - Break
- 5. Mixture modelling introduction & theory
- 6. DZmix application
- 7. DZnmf application
- Wrap-up

Module 7 Learning goals

- Understand the theory behind non-negative matrix factorization

- Understand how NMF can be used to identify unknown sediment sources.

- Understand how breakpoint analysis is used to determine the optimum factorization rank.

- Apply NMF using DZnmf.

Module 7 Outline

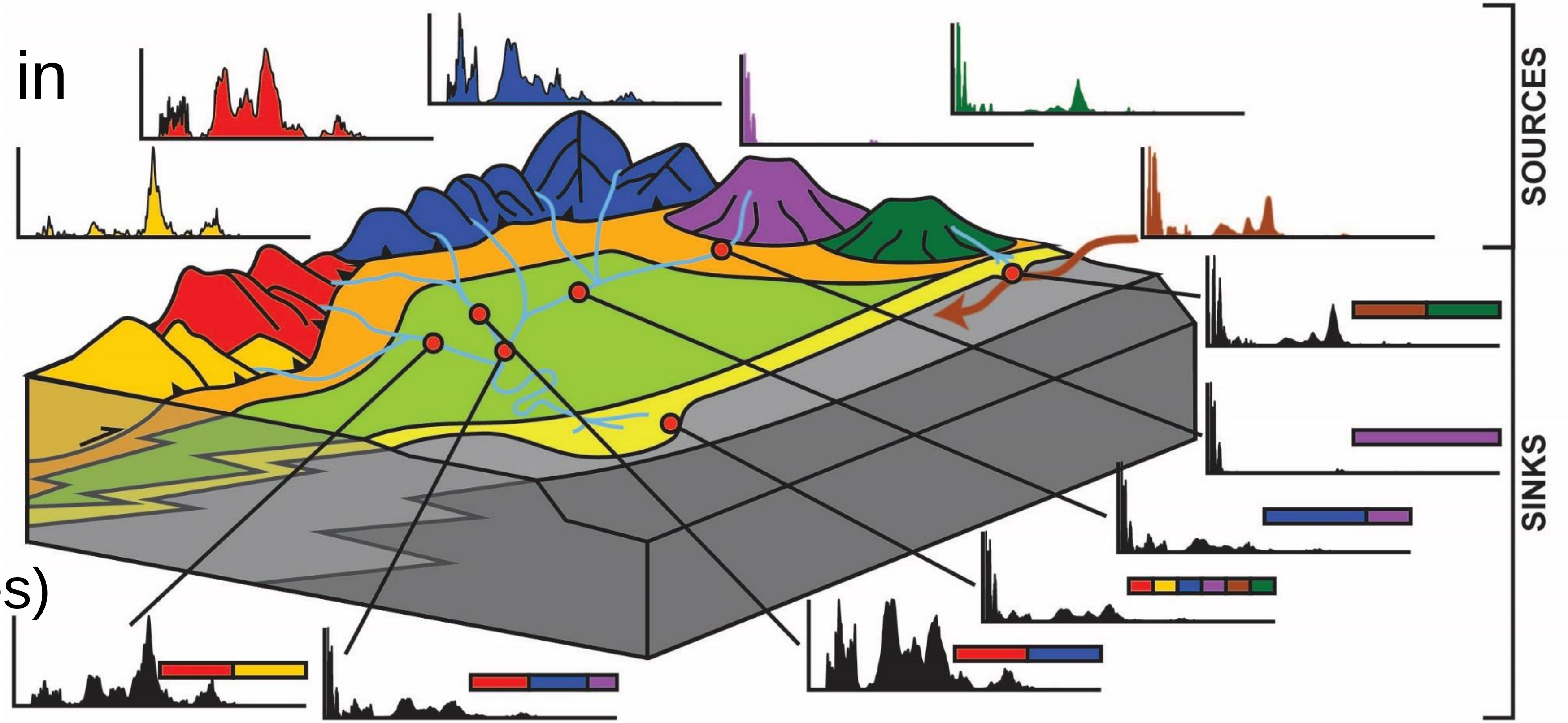
- Non-negative matrix factorization
    - NMF concept
    - NMF basics
    - Idealized example
        - Known and factorized age distributions
        - Known and factorized weights
- Determining the number of sources
- DZnmf
        - Factorizing a synthetic data set
        - Impact of the number of samples on factorization
        - Determining the optimum number of sources
        - NMF of an empirical data set.

Non-negative matrix factorization (NMF)

- "Bottom-up"
- Known sinks (Shown in Black)

  ↓ Factorization

- Unknown sources (Shown in Colors)
- Caveats
  - N(sinks)>> N (sources)
  - Sinks dissimilar
  - Sinks well characterized (large n)
  - Recycling is not always obvious



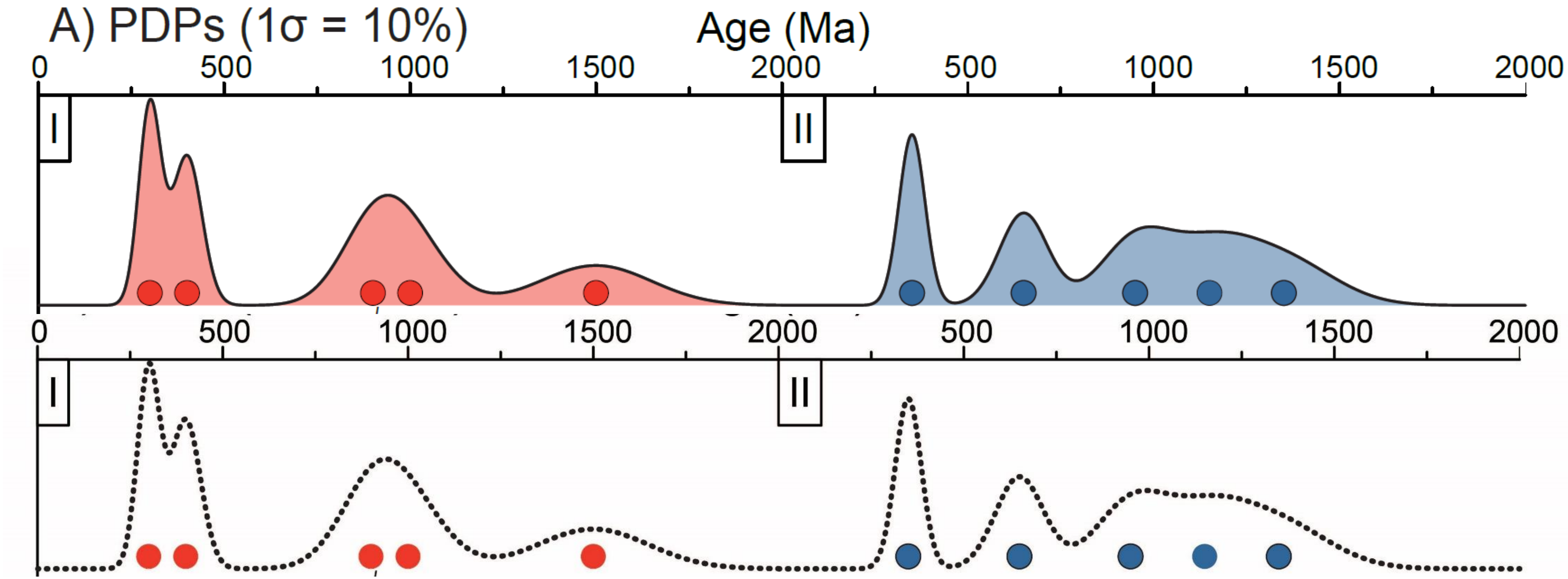
Inspired by Sharman and Johnstone (2017, EPSL)

Module 7 Outline

- Non-negative matrix factorization
  - NMF concept
  - NMF basics
  - Idealized example
    - Known and factorized age distributions
    - Known and factorized weights
- Determining the number of sources
- DZnmf
    - Factorizing a synthetic data set
    - Impact of the number of samples on factorization
    - Determining the optimum number of sources
    - NMF of an empirical data set.

Graphical representation

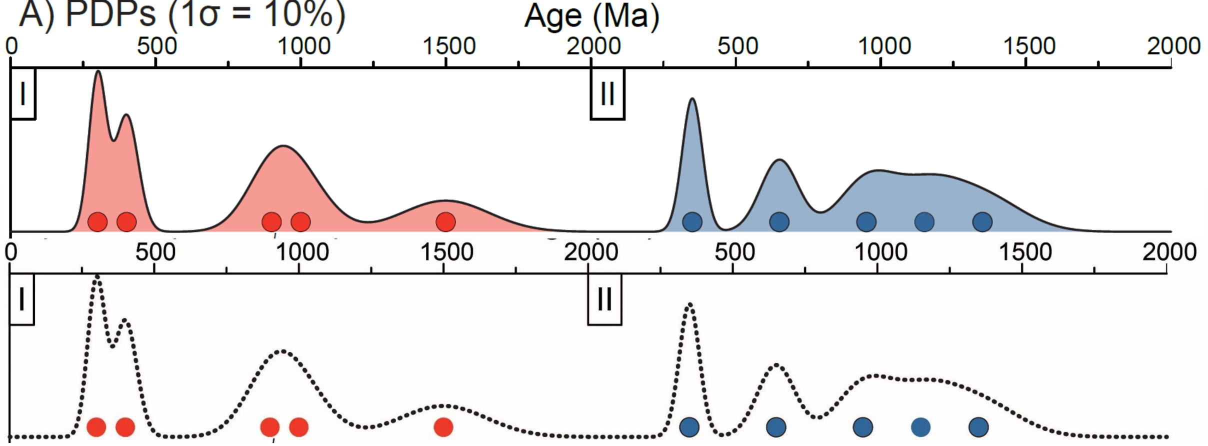
- Mixture distributions are matrices!
- Treat them as evenly spaced time series

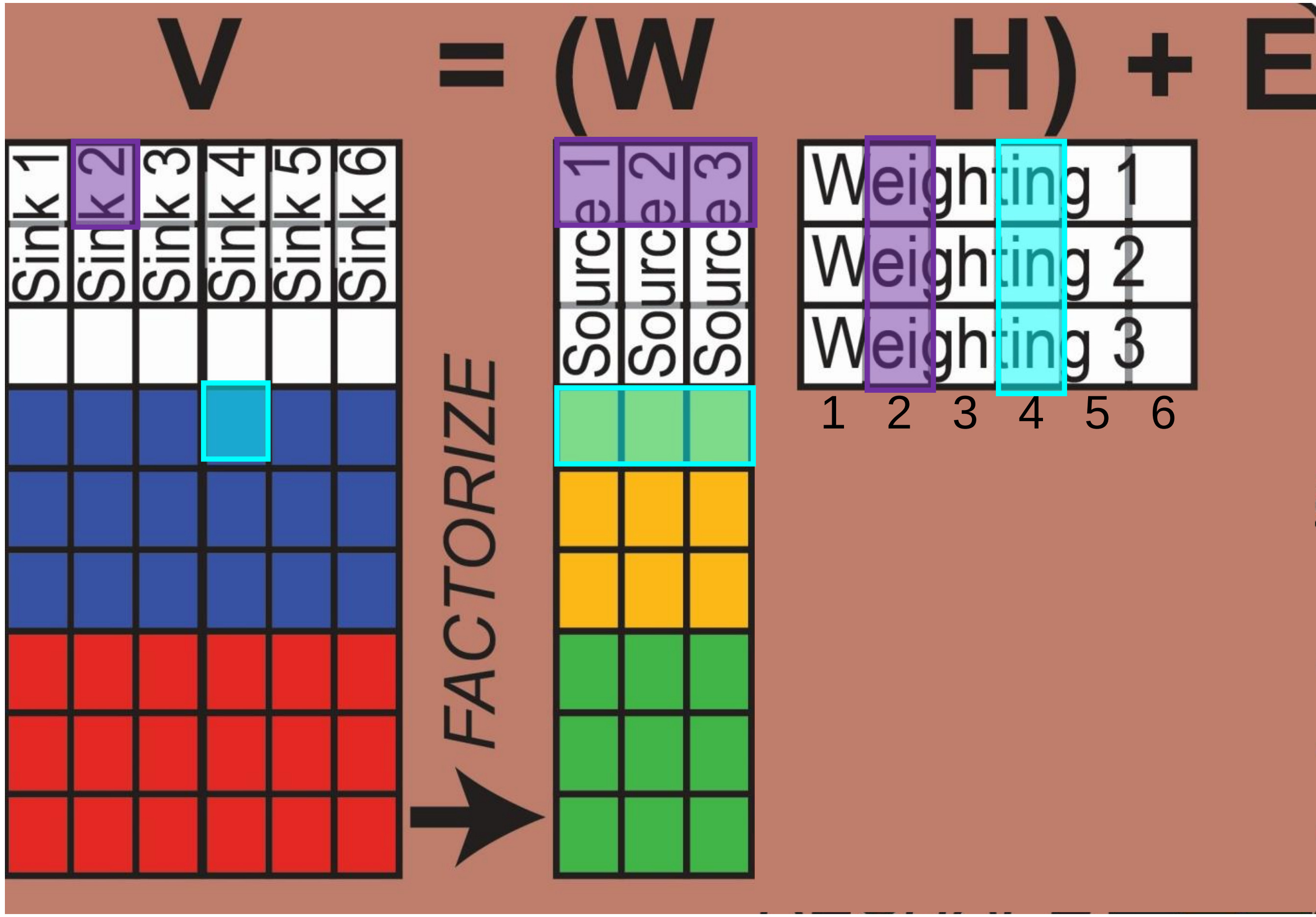


NMF Basics

- V: original non-negative data (m x n)
  - Samples in columns (n: detrital samples)
  - Features in rows (m: i.e., values of KDEs or PDPs)
- W: basis vectors (m x k)
  - k: number of sources (rank)
- H: weights (k x n)
  - (1,2) weighted elements of source 1, 2, 3
    - $(W_{1,1}H_{1,2} + W_{1,2}H_{2,2} + W_{1,3}H_{3,2})$
  - (4,4) weighted elements of source 1, 2, 3
    - $(W_{4,1}H_{1,4} + W_{4,2}H_{2,4} + W_{4,3}H_{3,4})$
  - etc

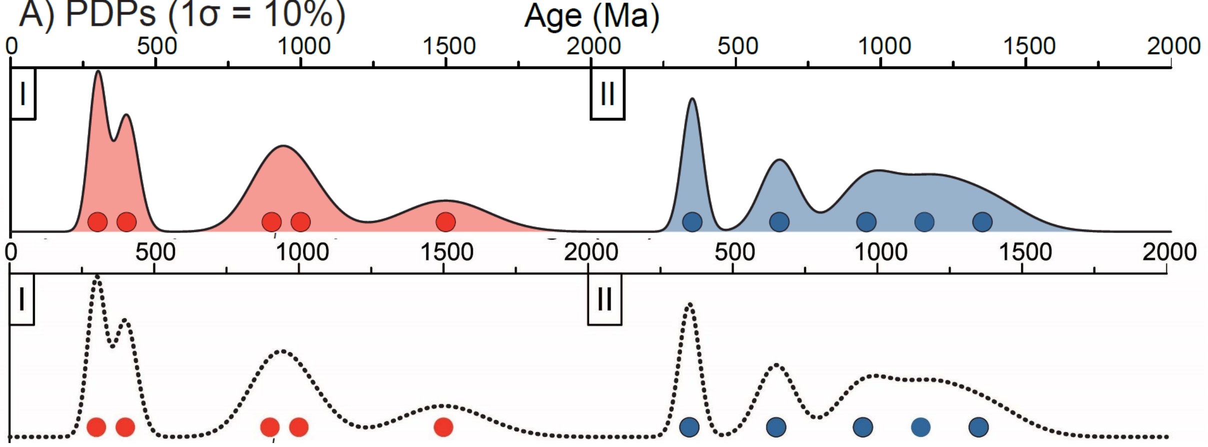
(Lee & Seung, 1999 & 2001)

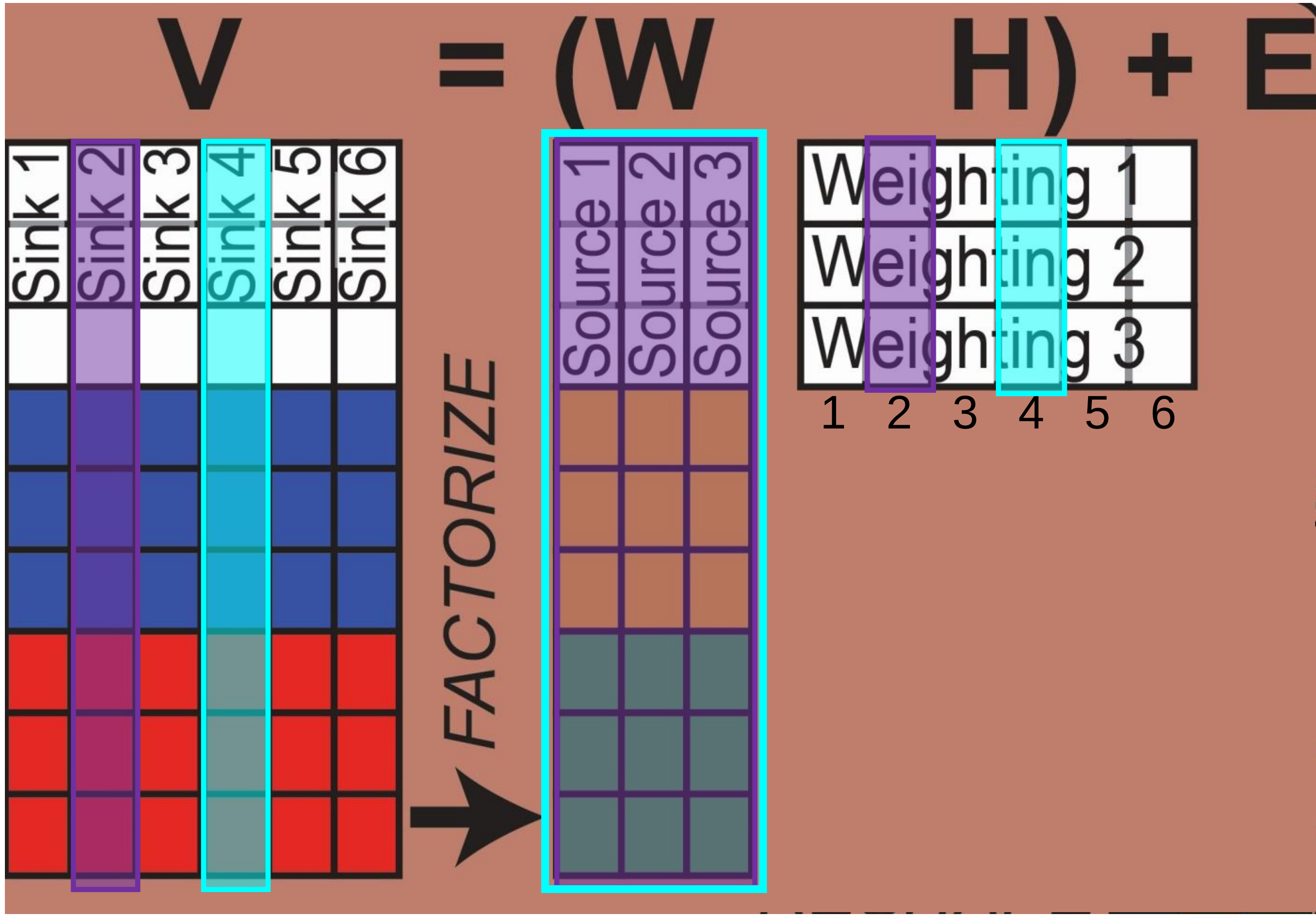




NMF Basics

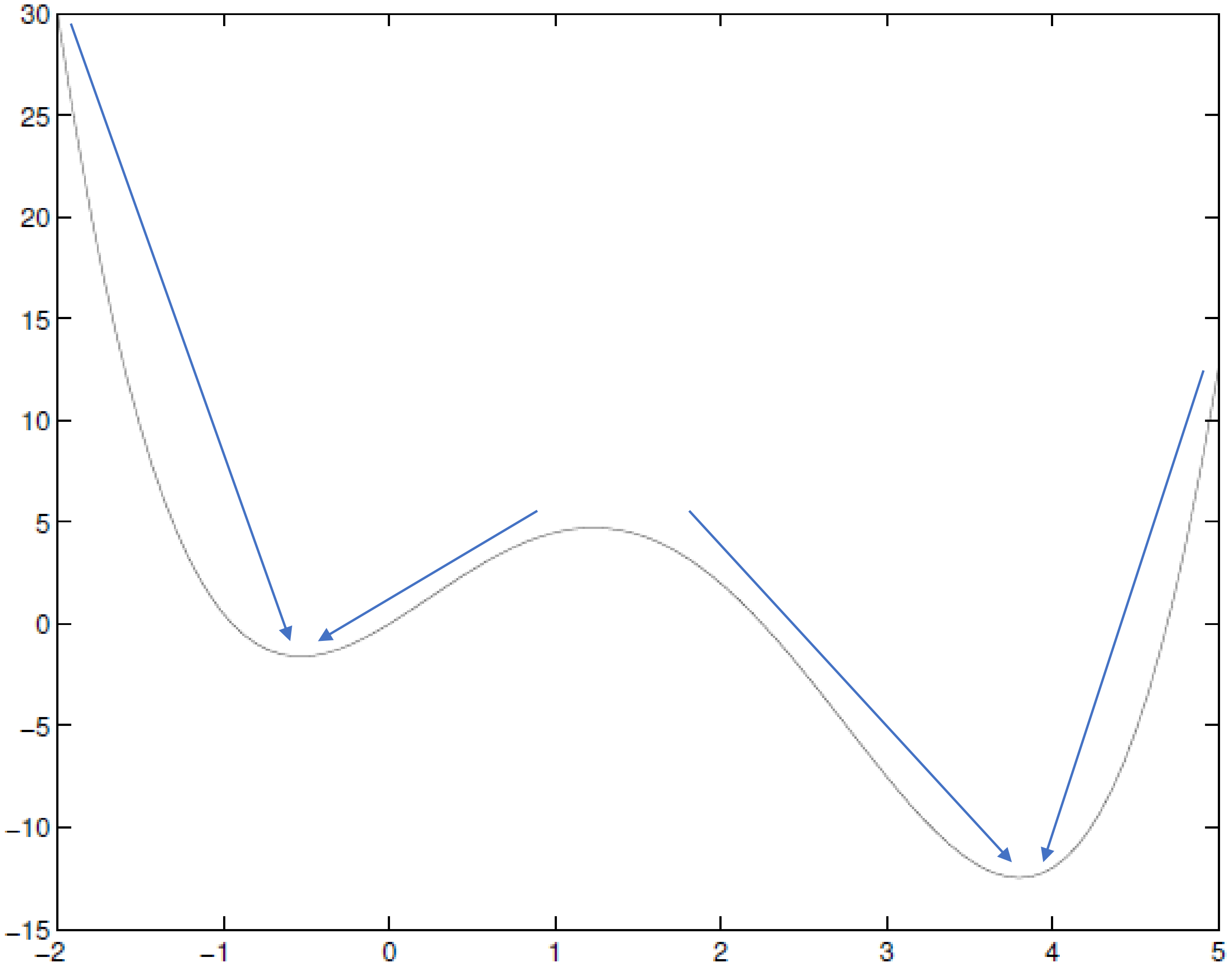
- i.e., columns of V are weighted sums of basis vectors (W)





NMF Basics

- CAVEATS
- NMF is non-convex
  - May find a local minimum
  - Sensitive to initial conditions
    - Initial conditions in DZnmf are randomized
  - MULTIPLE RUNS!



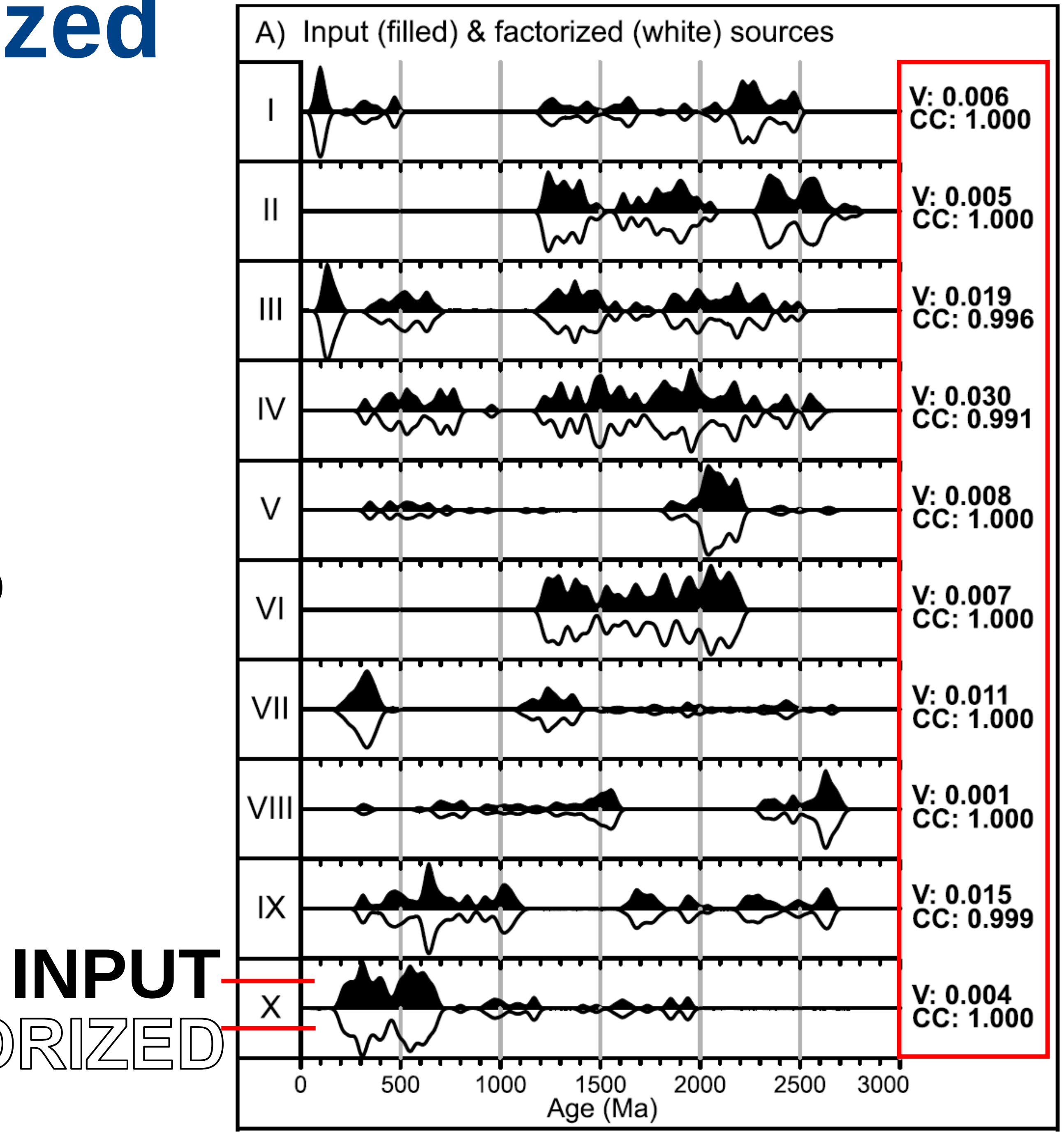
Module 7 Outline

- Non-negative matrix factorization
  - NMF concept
  - NMF basics
  - Idealized example
    - Known and factorized age distributions
    - Known and factorized weights
- Determining the number of sources
- DZnmf
    - Factorizing a synthetic data set
    - Impact of the number of samples on factorization
    - Determining the optimum number of sources
    - NMF of an empirical data set.

Known and factorized age distributions

$$V = \boxed{W} H + E$$

- Synthetic sources from Sundell and Saylor (2017)
- KDEs 20 Myr bandwidth
- **Input** sources randomly mixed into 40 sink samples
  - Factorized with no training or supervision
- **Cross-correlation and Kuiper V indicate nearly perfect matches**

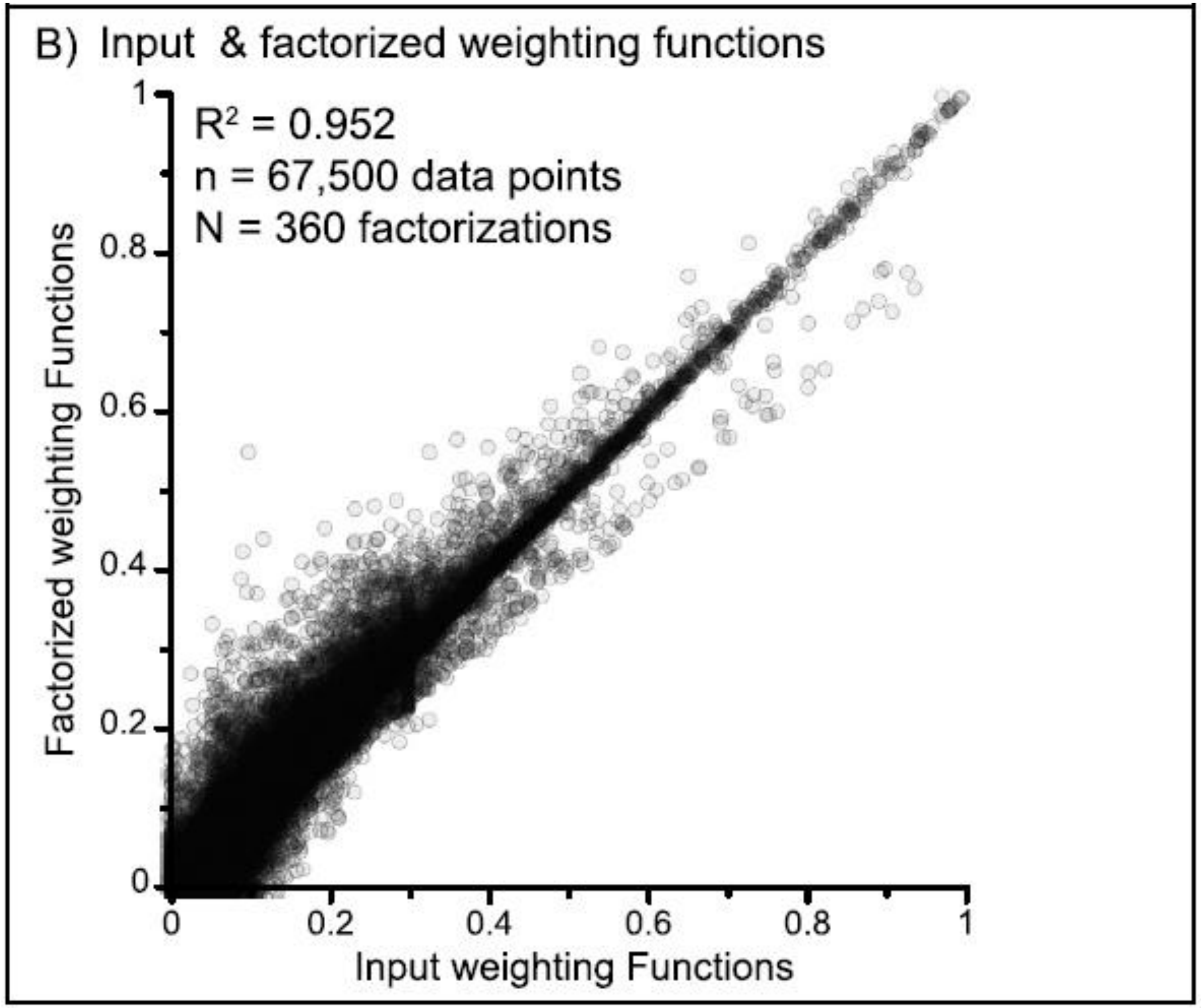


Saylor et al. (2019, EPSL)

Known and factorized weights

$$V = W\boxed{H} + E$$

- Comparison of input and factorized weighting functions
- $R^2 = 0.95$



Saylor et al. (2019, EPSL)

Module 7 Outline

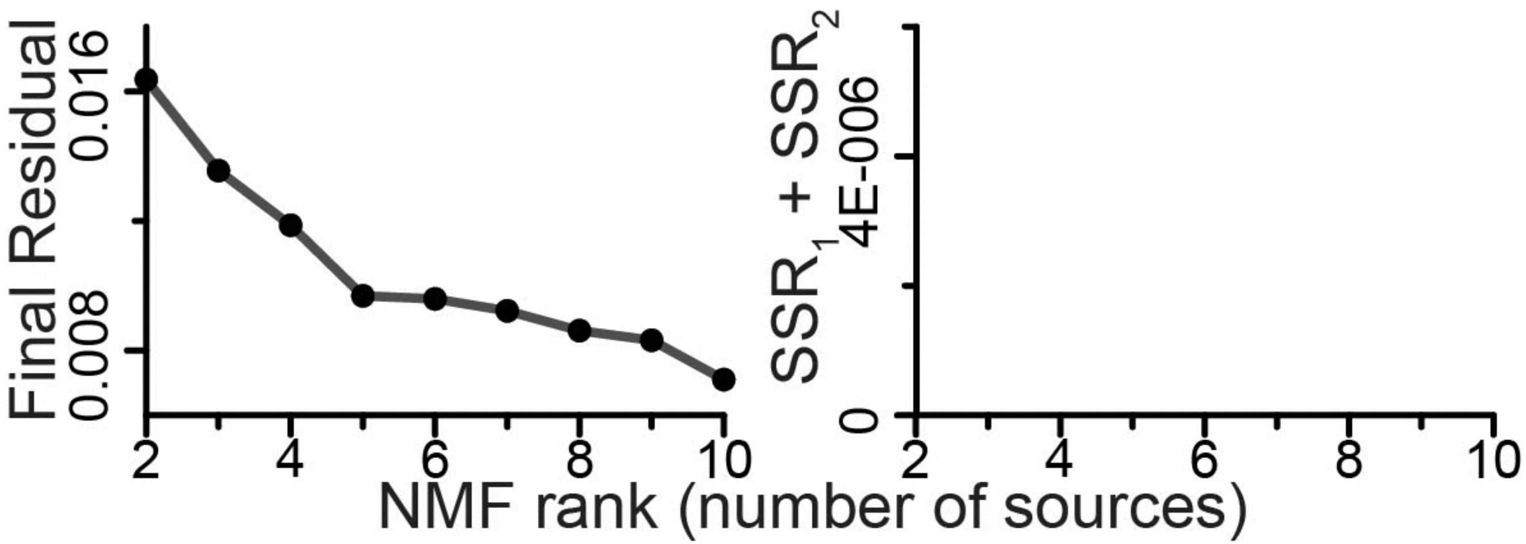
- Non-negative matrix factorization
  - NMF concept
  - NMF basics
  - Idealized example
    - Known and factorized age distributions
    - Known and factorized weights
- Determining the number of sources
- DZnmf
    - Factorizing a synthetic data set
    - Impact of the number of samples on factorization
    - Determining the optimum number of sources
    - NMF of an empirical data set.

Determining the number of sources

$$SSR_1 = \sum_{r=2}^{r=x_b} \left(R_r - f\left(x_r\right)\right)^2$$

and

$$SSR_2 = \sum_{r=x_b}^{r=n} \left(R_r - \left(g(x_r)\right)\right)^2$$



Saylor et al. (2019, EPSL)

- R_r = final residual
- f(x) and g(x) = predicted value for linear fit
- CAVEATS
  - The breakpoint is dependent on the ranks tested (Test to a higher rank)

Module 7 Outline

- Non-negative matrix factorization
  - NMF concept
  - NMF basics
  - Idealized example
    - Known and factorized age distributions
    - Known and factorized weights
- Determining the number of sources
- DZnmf: see Step-by-Step guide for instructions
    - Factorizing a synthetic data set
    - Impact of the number of samples on factorization
    - Determining the optimum number of sources
    - NMF of an empirical data set.

Optimization

- 1. Initialize the entries in **W** and **H** with random positive
- values
- 2. Update **W**
- 3. Update **H**
- 4. Iterate steps 2 and 3

Input controls on results (W)

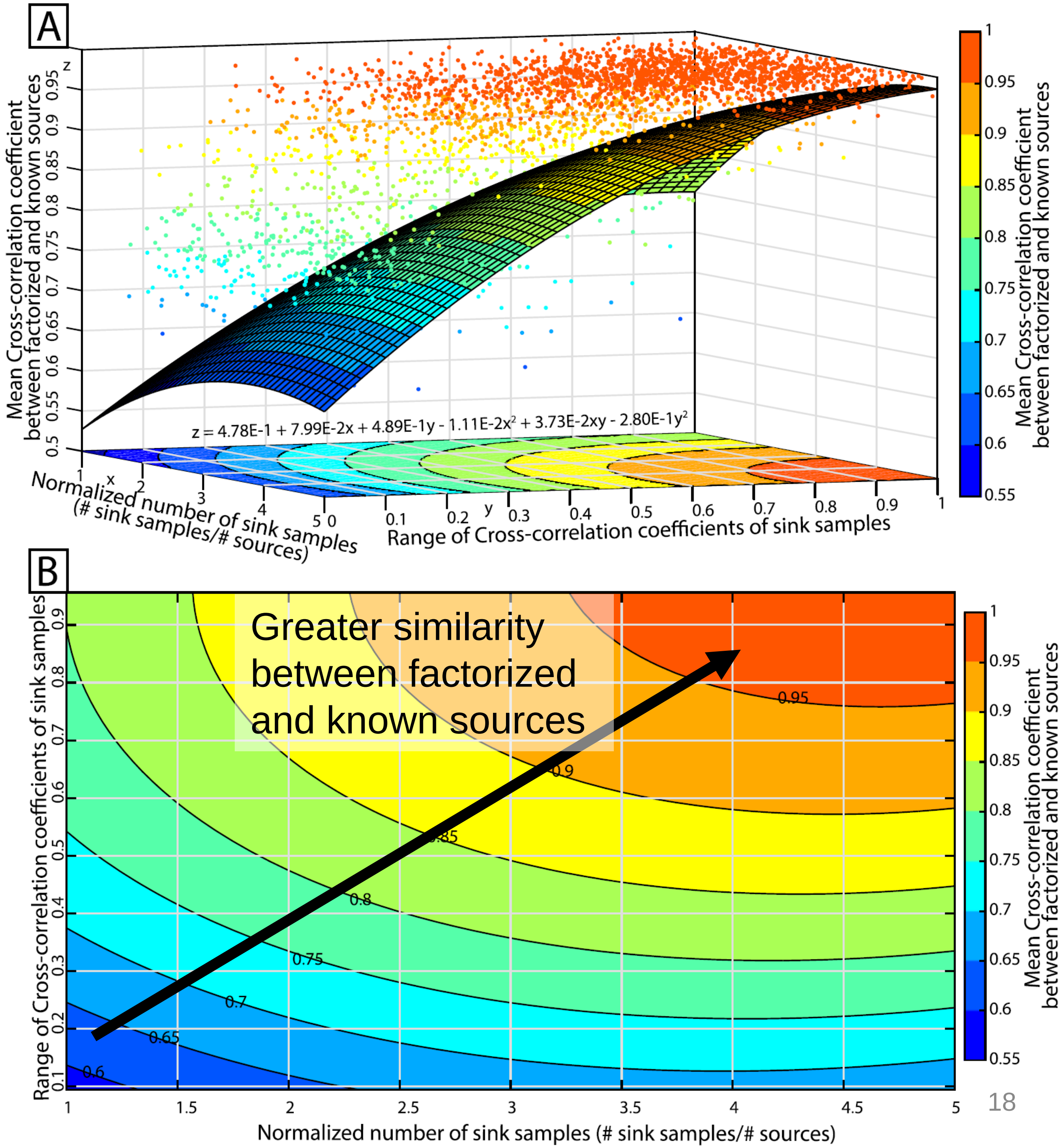
- Greater dissimilarity between input sinks

&

- More sink samples

Results in

- Closer match between factorized and known sources



Saylor et al. (2019, EPSL)

Input controls on results (W)

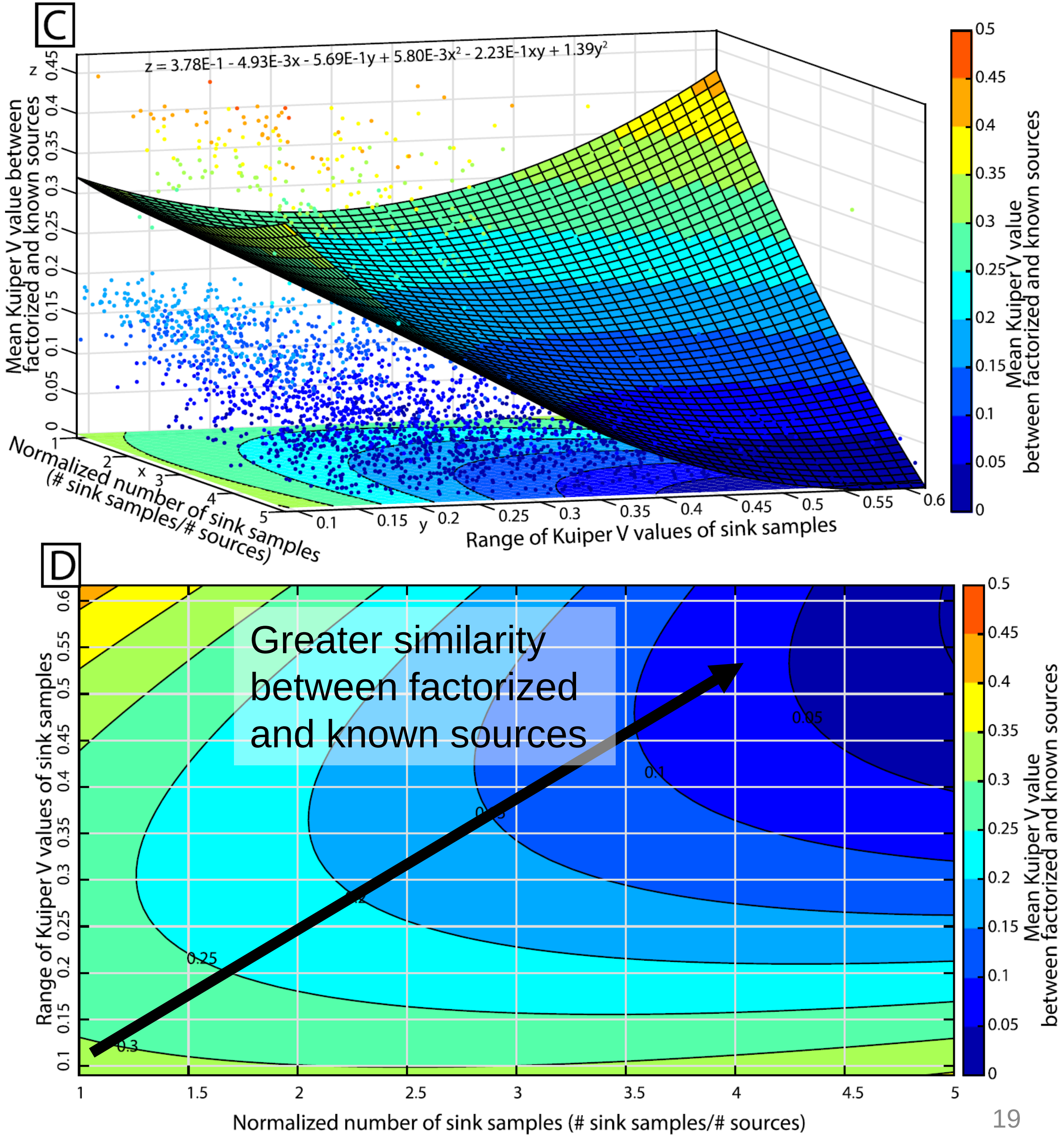
- Greater dissimilarity between input sinks

&

- More sink samples

Results in

- Closer match between factorized and known sources



Saylor et al. (2019, EPSL)

Input controls on results (W)

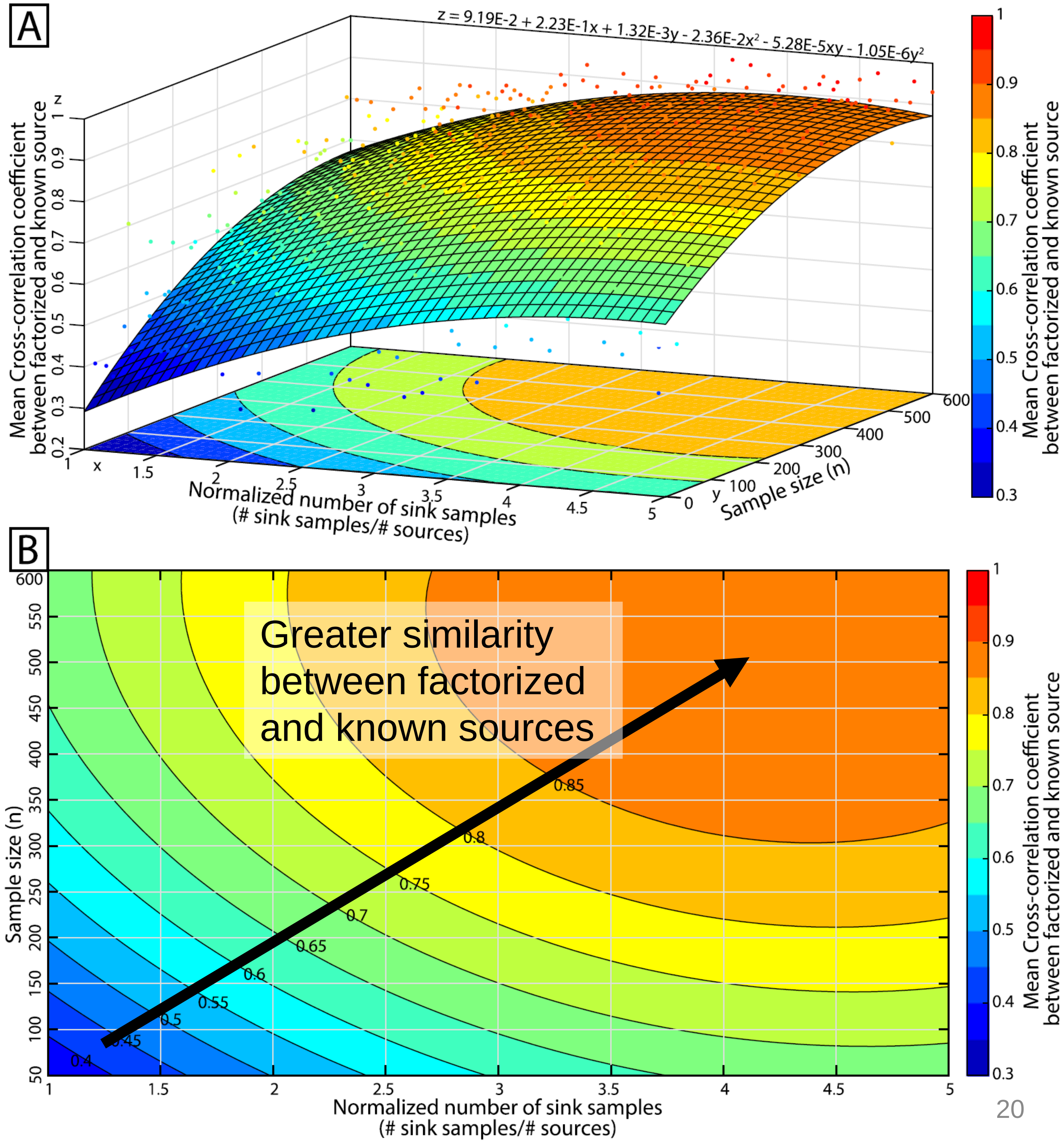
- Greater sink size

&

- More sink samples

Results in

- Closer match between factorized and known sources



Saylor et al. (2019, EPSL)

Input controls on results (W)

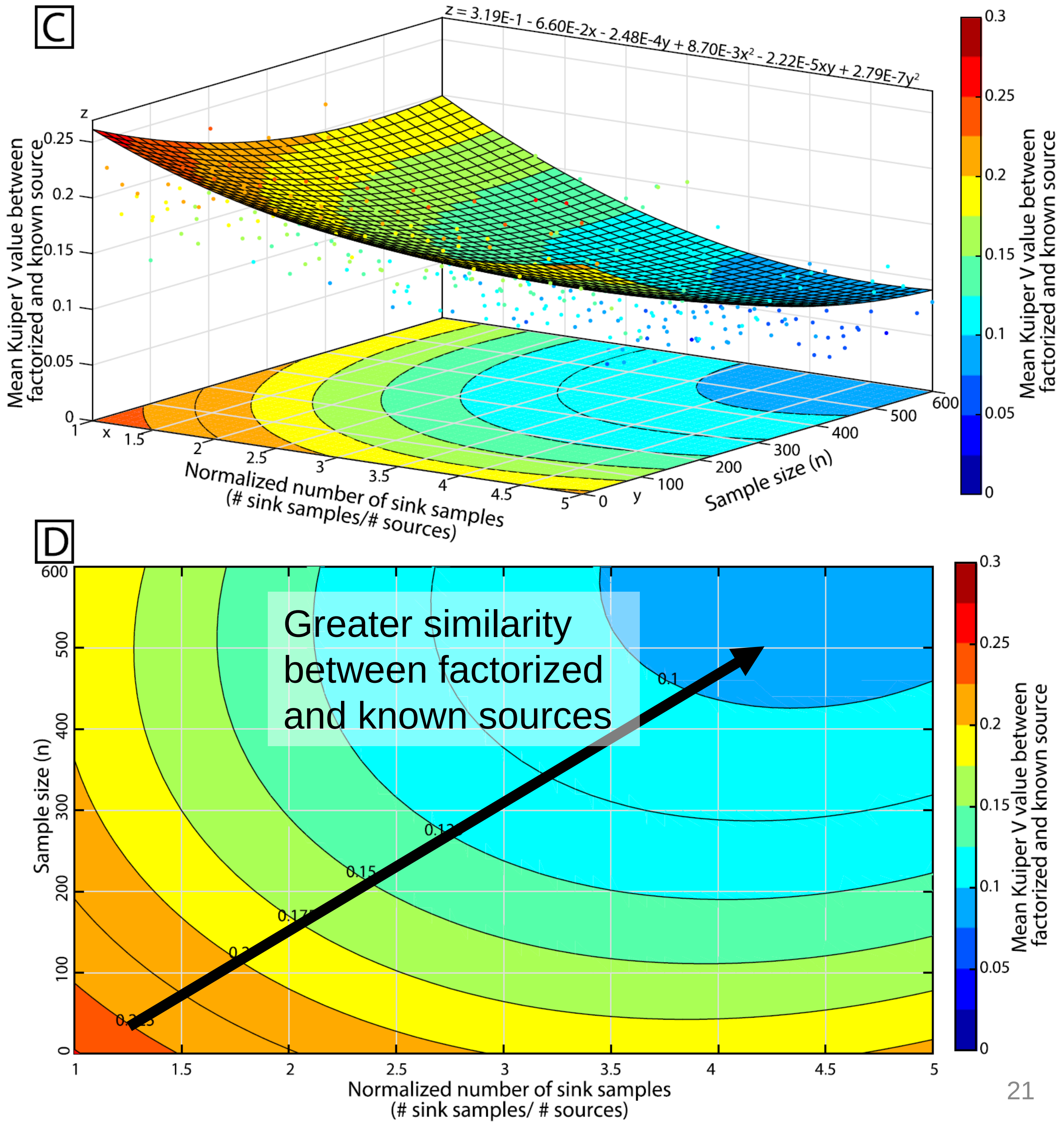
- Greater sink size

  &

- More sink samples

  Results in

- Closer match between factorized and known sources



Saylor et al. (2019, EPSL)

Input controls on results (H)

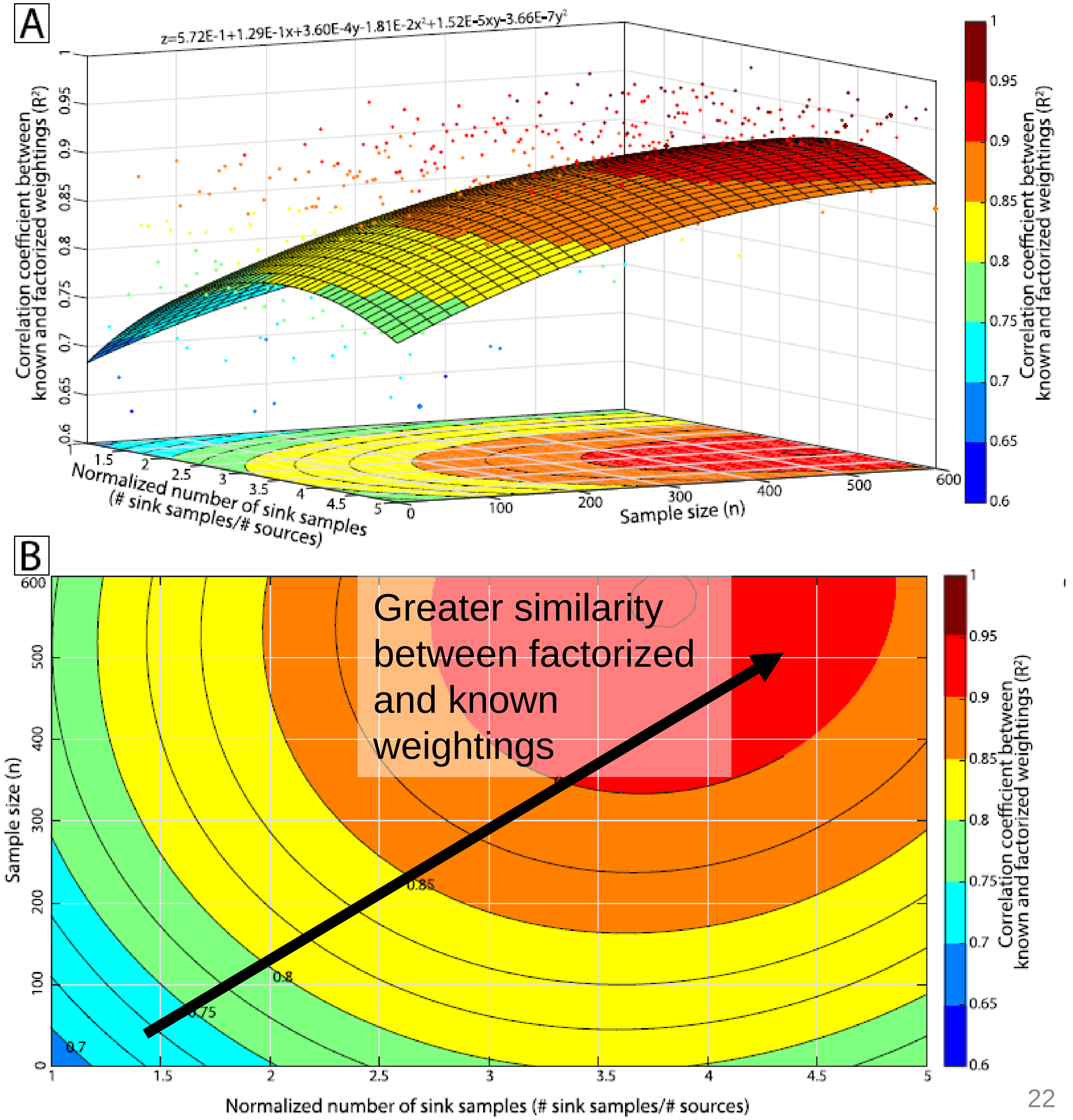
- Greater sink size

&

- More sink samples

Results in

- Closer match between factorized and known sources



Saylor et al. (2019, EPSL)

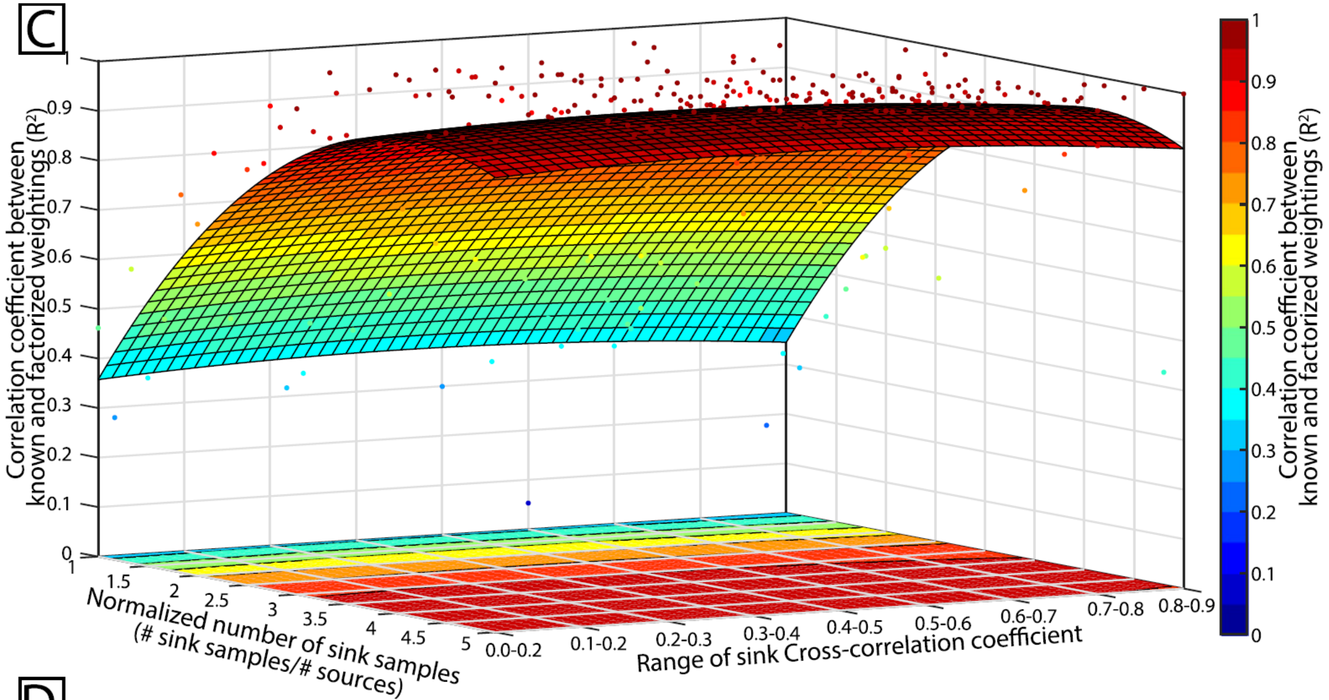
Input controls on results (H)

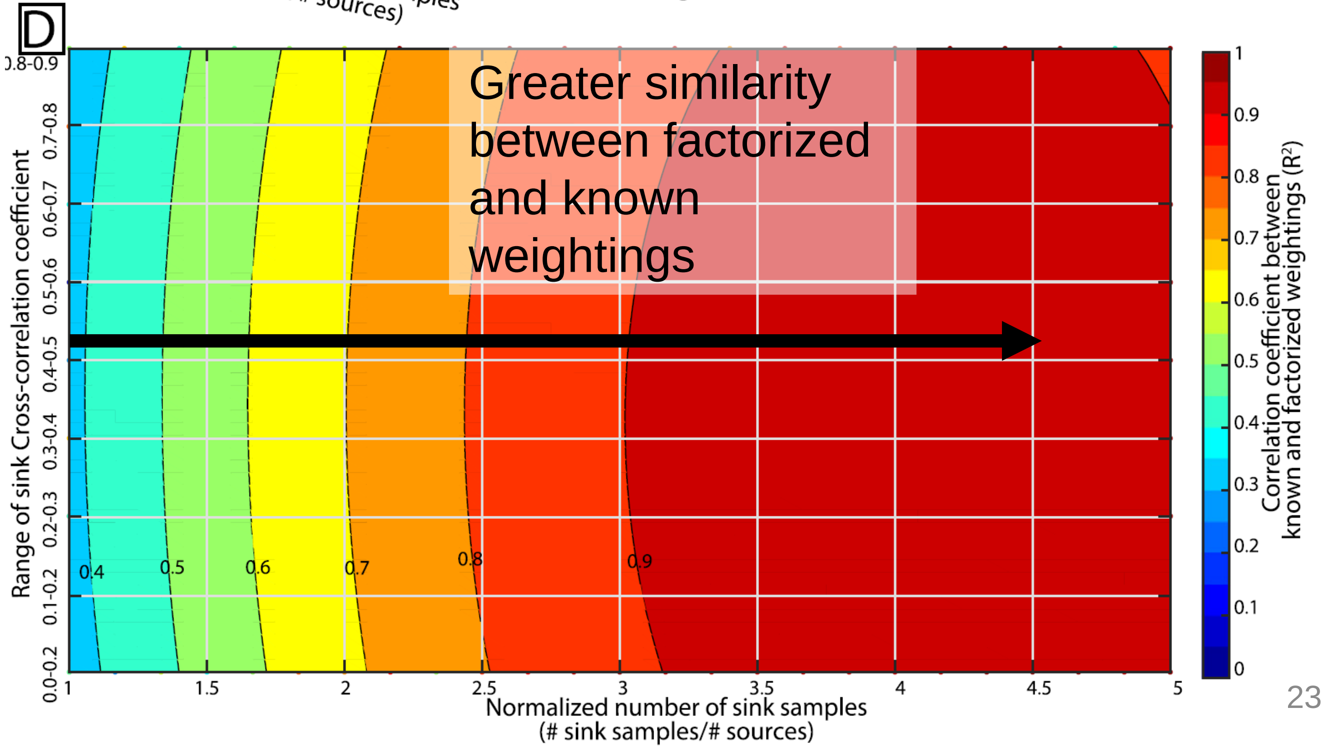
- More sink samples

  Results in

- Closer match between factorized and known sources

- Greater dissimilarity between sink samples does not affect similarity of factorized and known weights.





Saylor et al. (2019, EPSL)