Parallel Hierarchical Clustering using Rank-Two Nonnegative Matrix Factorization

Lawton Manning[1] Grey Ballard[1]
Ramakrishnan Kannan[2] Haesun Park[3]

[1]Wake Forest University

[2]Oak Ridge National Laboratory

[3]Georgia Institute of Technology

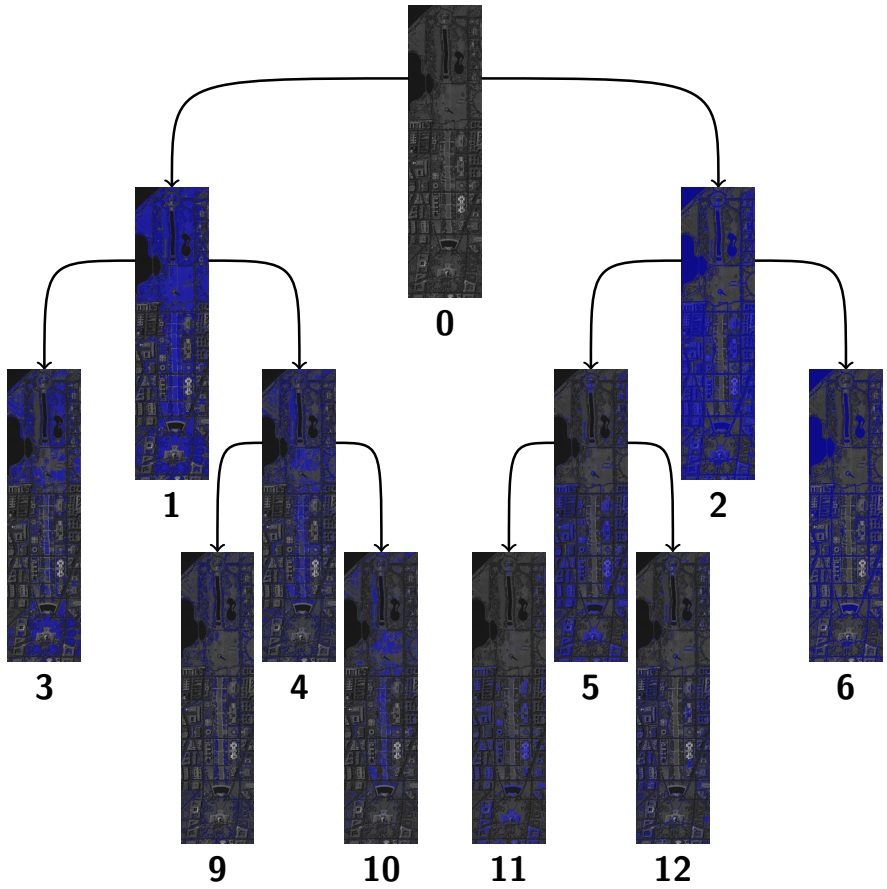
HiPC 2020

Nonnegative Matrix Factorization (NMF)

- $A \approx WH^T$
- approximate A (features $\times$ samples) into W (features $\times k$) and H (samples $\times k$)
- nonnegativity gives interpretability of W and H as clusters and cluster membership, respectively
- applications
  - text document clustering using bag of words or TF-IDF matrices
  - hyperspectral imaging segmentation

Hierarchical NMF

- repeatedly use NMF with $k = 2$ to create a hierarchical tree of clusters
- application: hyperspectral imaging



Solving NMF

- constrained optimization:

$$\min_{W,H \geqslant 0} \|A - WH^T\|_2$$

- alternating update
  - fix H and solve the NNLS for W exactly
  - alternate and repeat until convergence
- NNLS Solvers
  - Block Principal Pivoting (BPP)
  - ??? Other Method ???

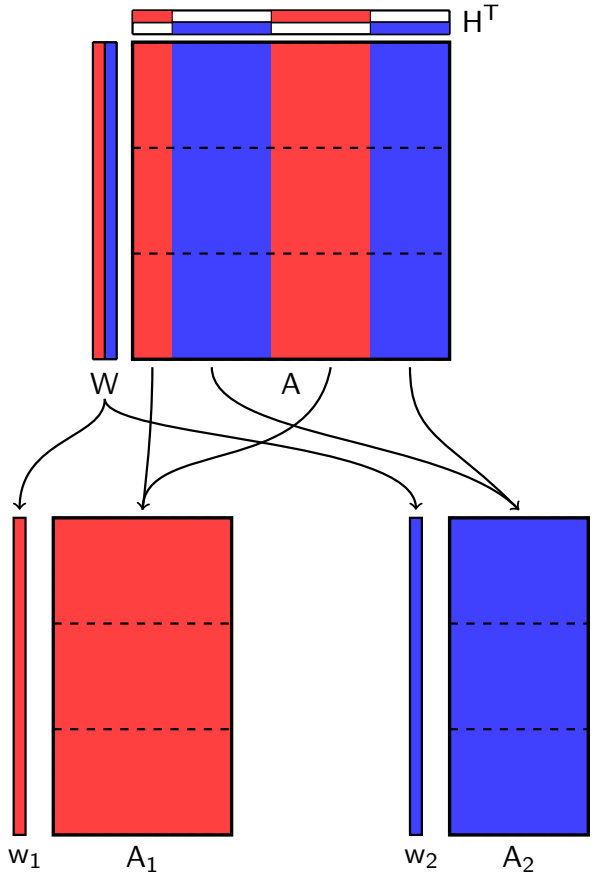
Rank-2 NMF (R2NMF)

- when $k = 2$, BPP can be solved quickly as the size of the active set is 4
- Active Sets for R2NMF (row by row)
  - both columns nonnegative
  - only left column nonnegative
  - only right column nonnegative
  - both columns negative

Parallel R2NMF

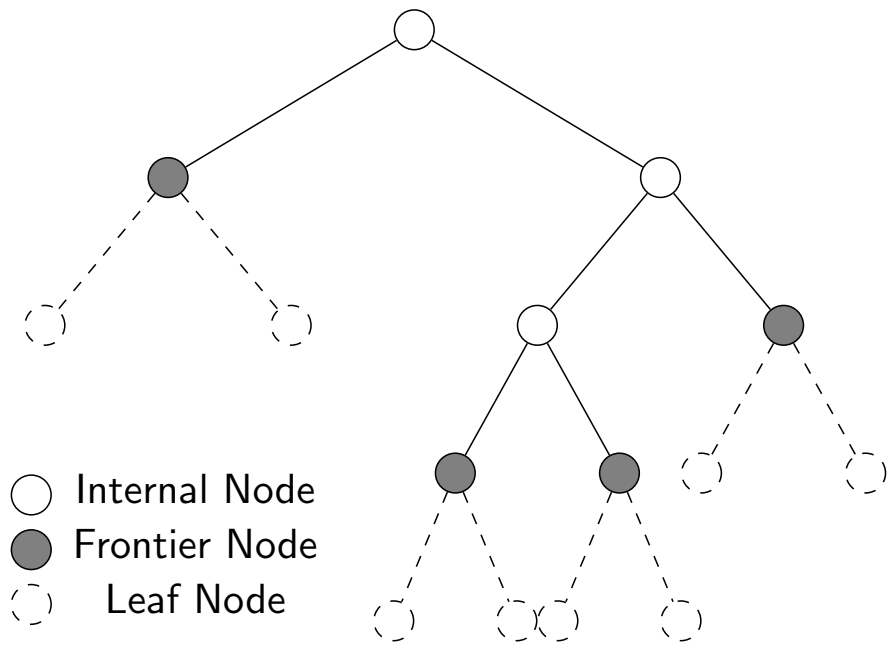
Splitting with R2NMF

- use columns of H^T to split the columns of A into two submatrices
- assign the corresponding columns of W to the submatrices
- repeat NMF and split on submatrices



HierNMF

- call R2NMF on root node to split into two children
- use the power iteration to select children that least approximate rank-1 matrices
- repeat until maximum number of nodes is reached or all child matrices are rank-1



Parallel HierNMF

FIXME: Need data distribution figure

Data Sets

- DC-HYDICE
  - Hyperspectral Digital Imagery Collection Experiment (HYDICE) of the National Mall in Washington, DC
- SIIM-ISIC
  - Society for Imaging Informatics in Medicine - International Skin Imaging Colloboration image classification of melanoma images
- Synthetic Image classification
  - smaller image classification dataset which has the same aspect ratio as SIIM-ISIC but small enough to fit in memory on one machine