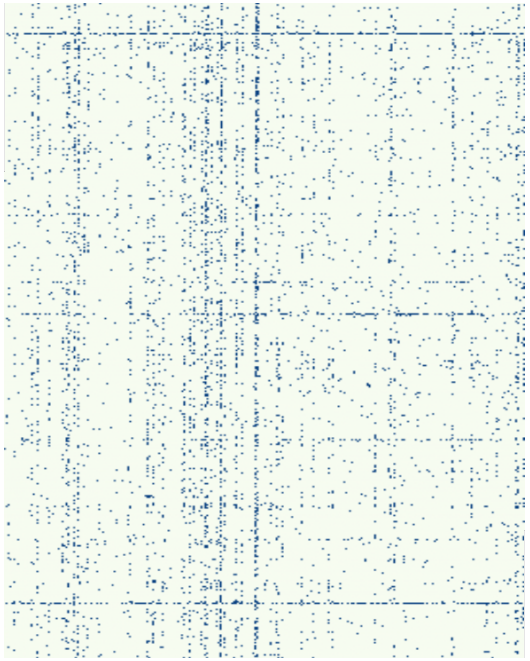
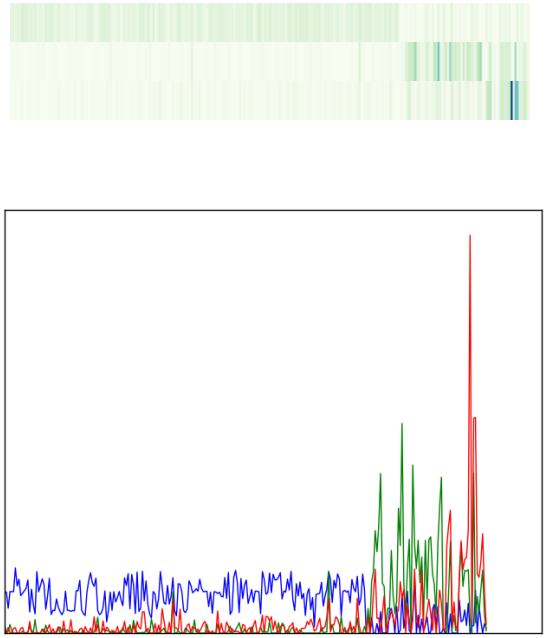
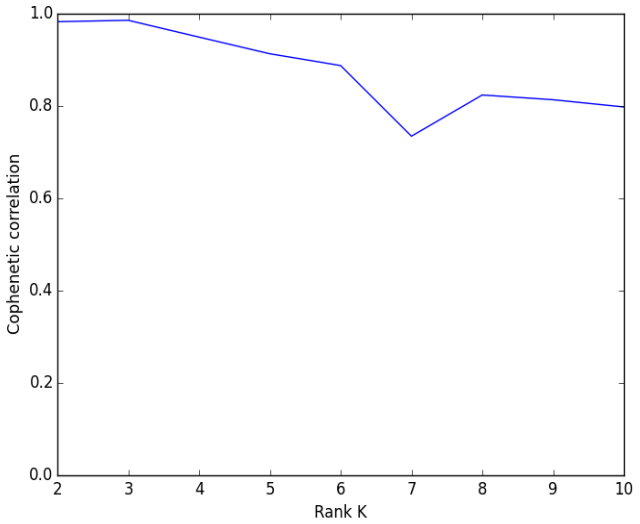
COAD

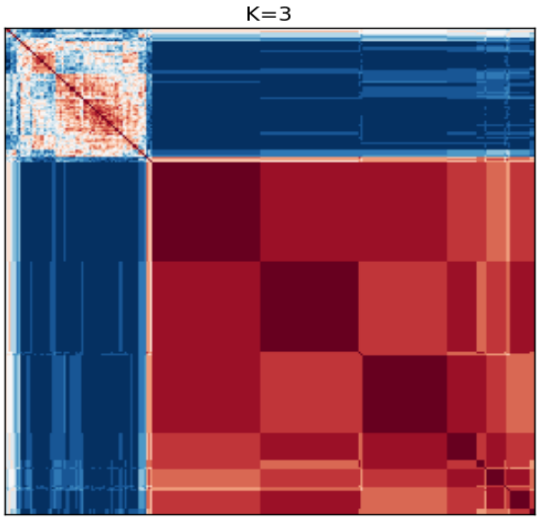
Best result: k = 3

A ≈ W H





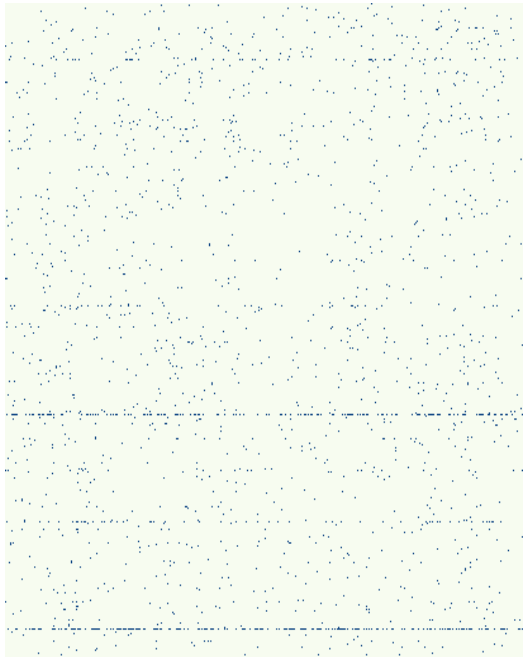


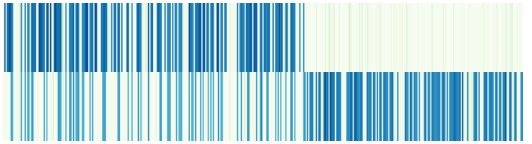


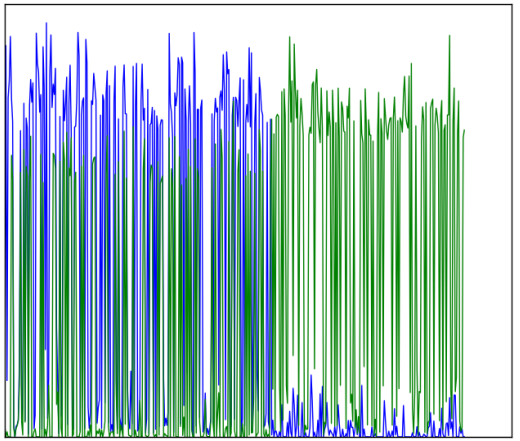
KIRC

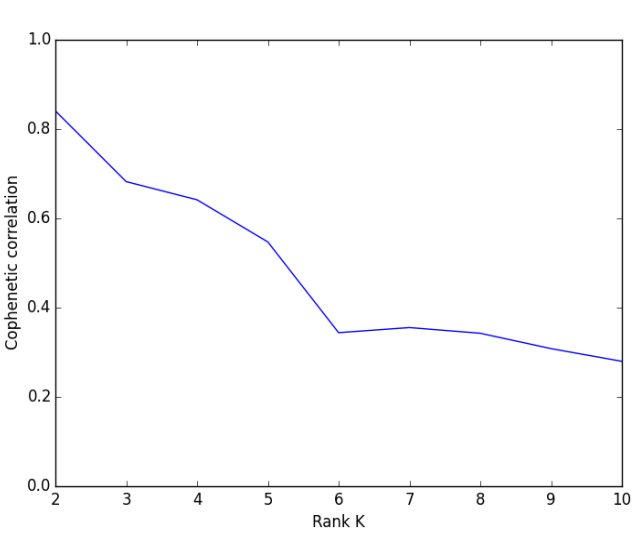
Best result: k = 2

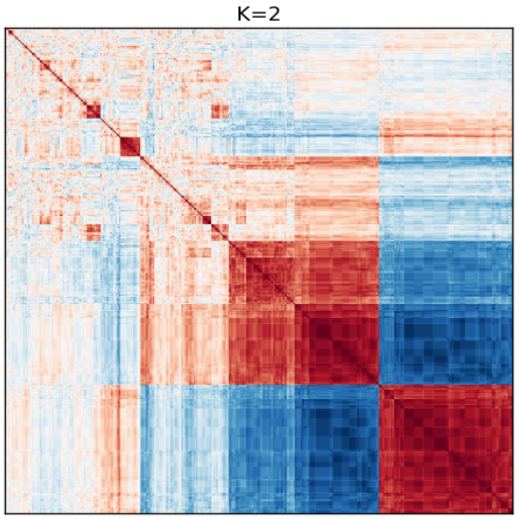
A ≈ W H







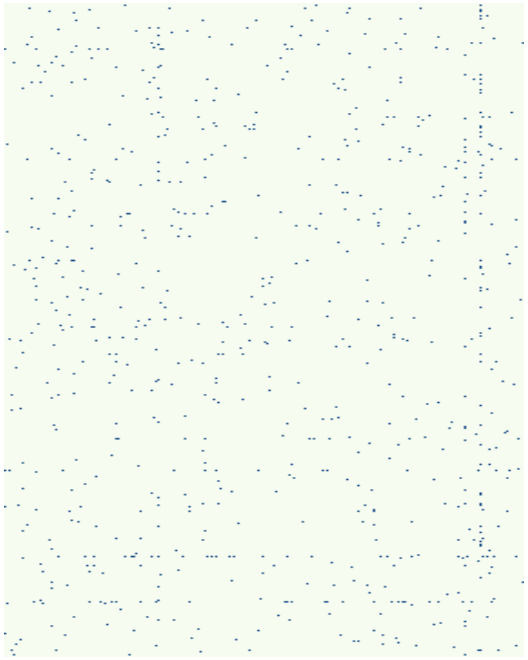


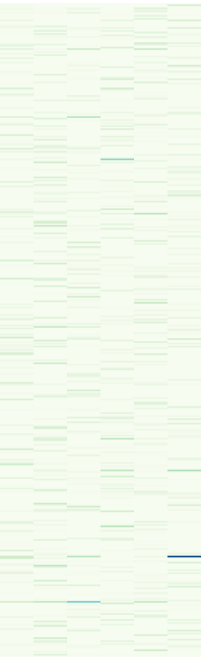


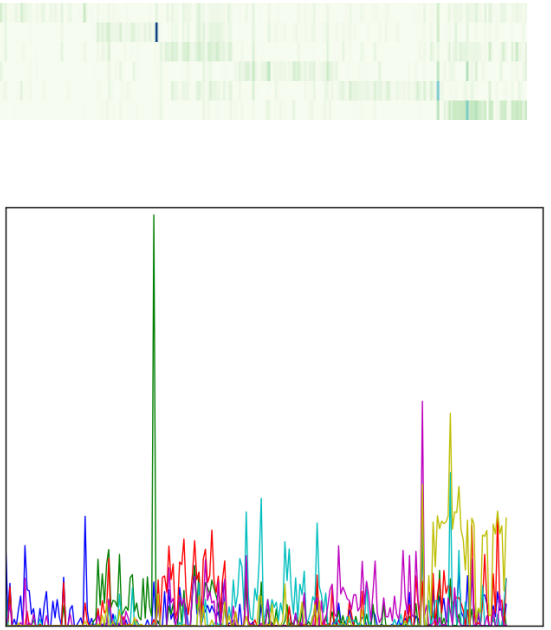
PRAD

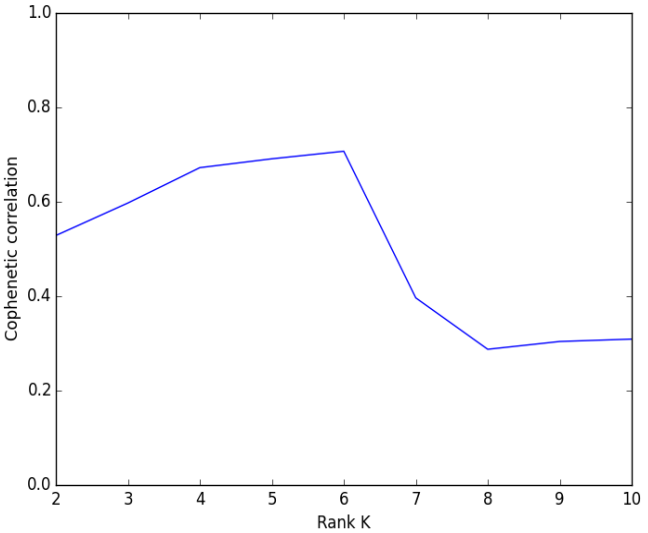
Best result: k = 6

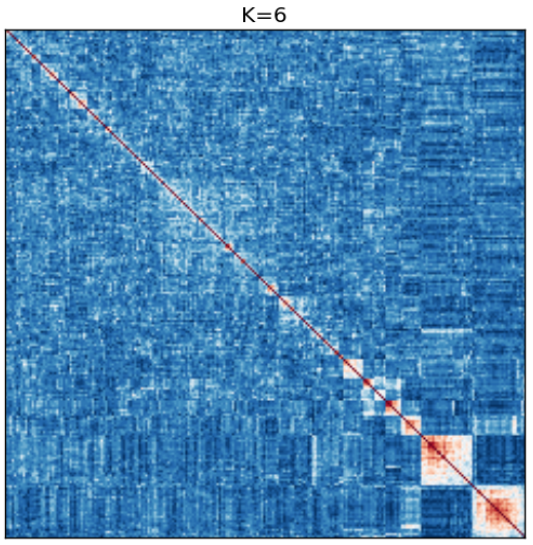
A ≈ W H







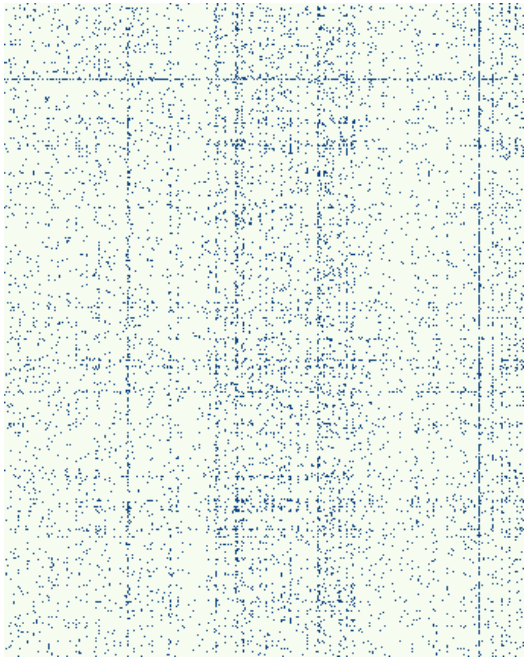


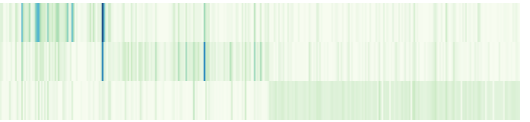


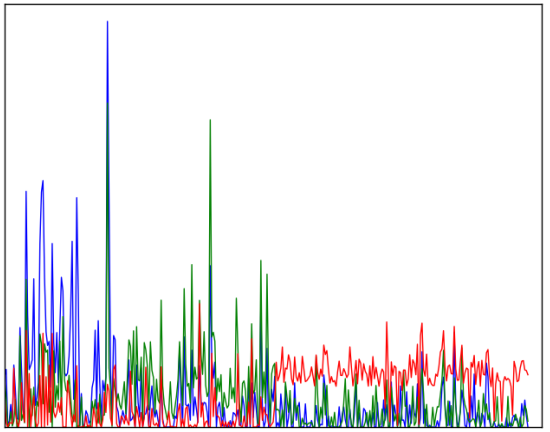
SKCM

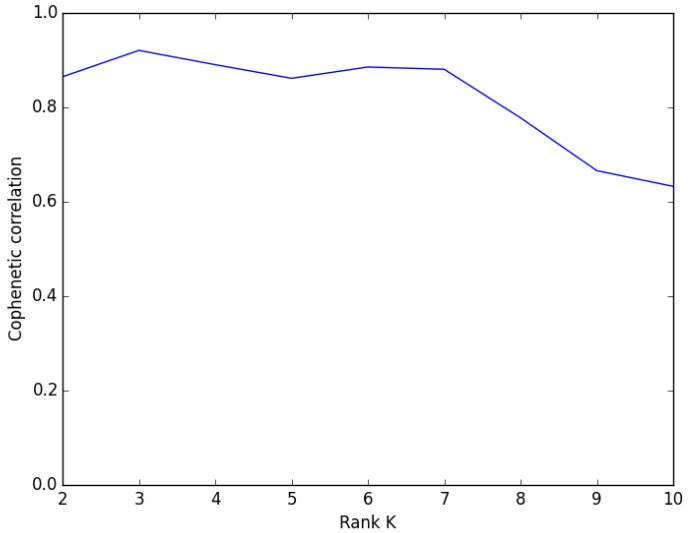
Best result: k = 3

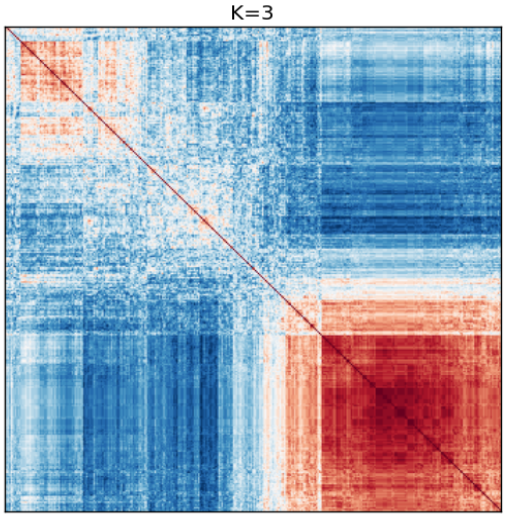
A ≈ W H







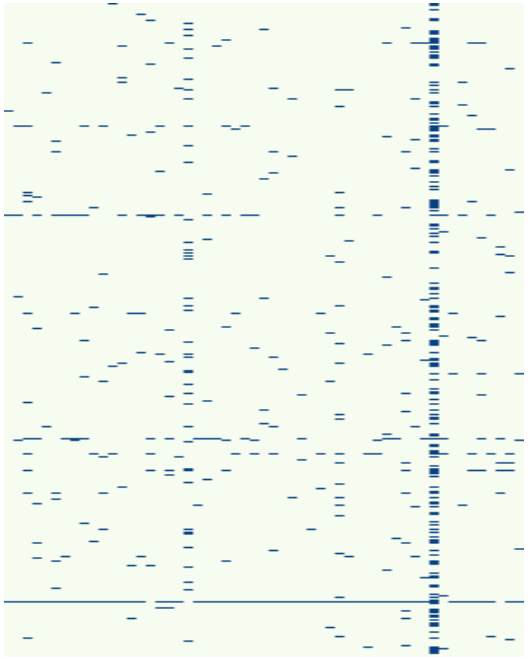


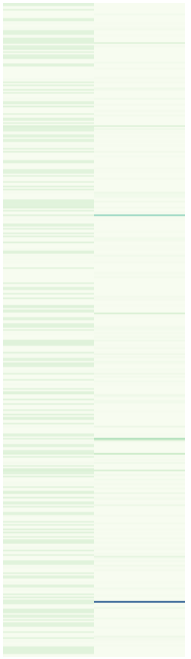


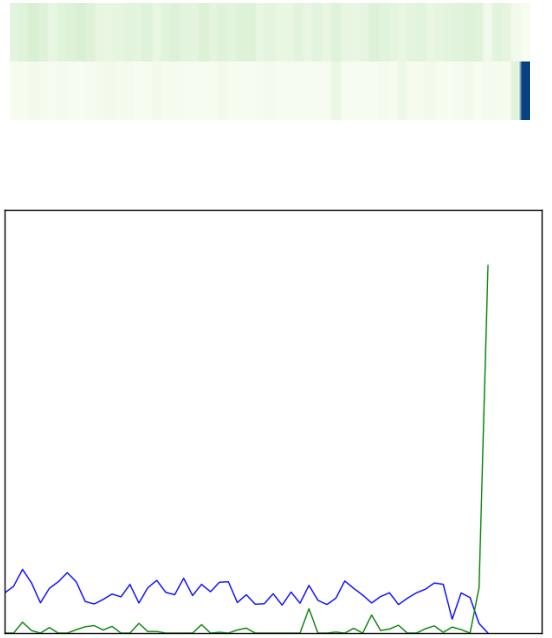
UCS

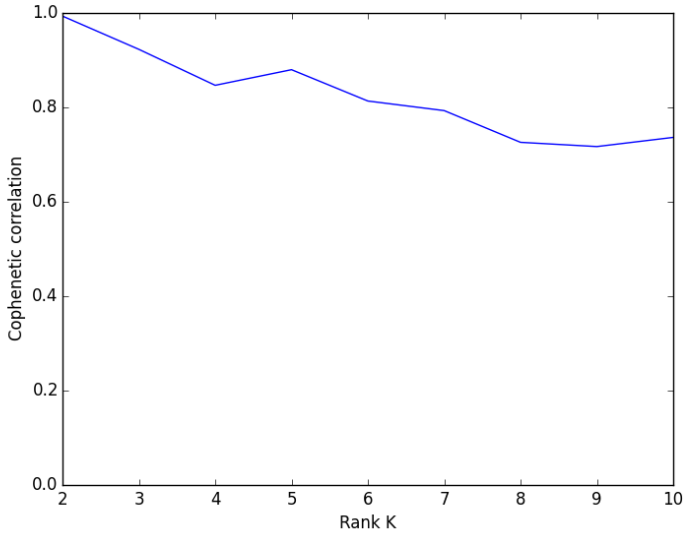
Best result: k = 2

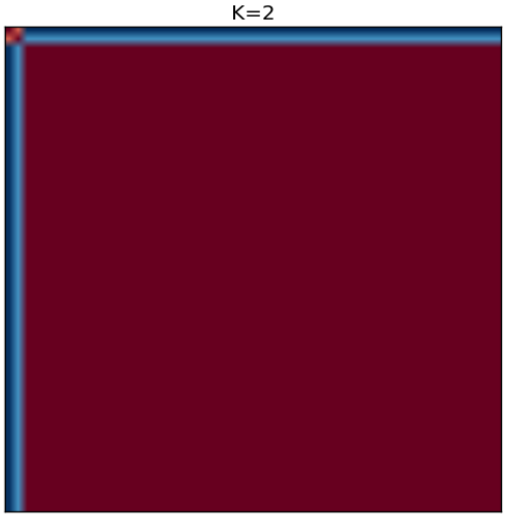
A ≈ W H











Dimension Reduction

Training Logistic Regression

- Trained on entire data set, then tested on that same data set.
  - Result: 31.2% accuracy
  - If convert matrix to one-hot: 98% accuracy! Presumed overfitting, verified by splitting data into test & training sets. Showed 0% accuracy

Feature selection

Ran sparse PCA algorithm to select 100 genes. Inspired by Hsu and Huang's *Sparse principal component analysis in cancer research. Sparse PCA offers* " powerful data reduction functionality and incorporating the sparseness model for variable selection"

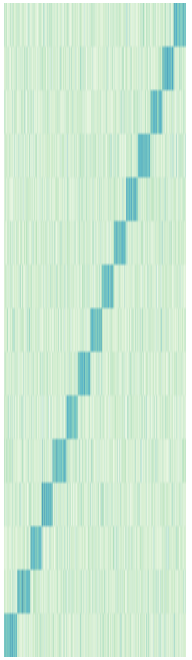
Logistic regression results with sparse PCA – inconclusive

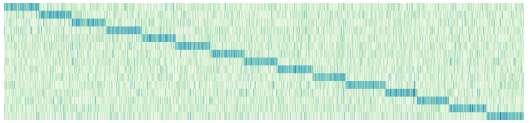
- 64.3% accuracy on test data, but did not remain stable across runs
- Approximately 30% accuracy with 10% training set for each cancer.
  - Still not very good, considering random guessing would provide 25%
- Requires further investigation / development to get repeatable results

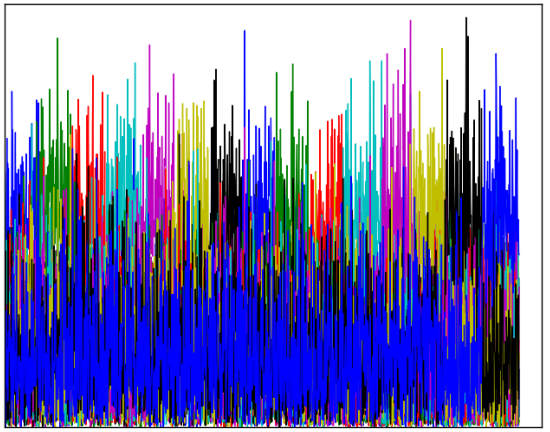
All cancer types

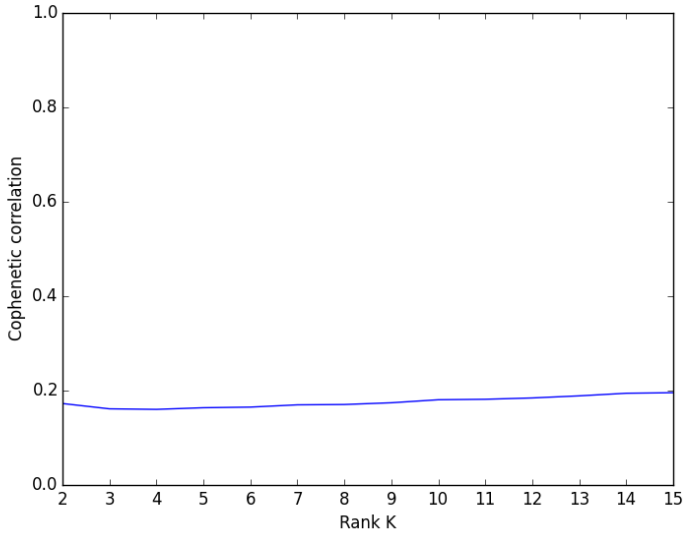
Best result: k = 15

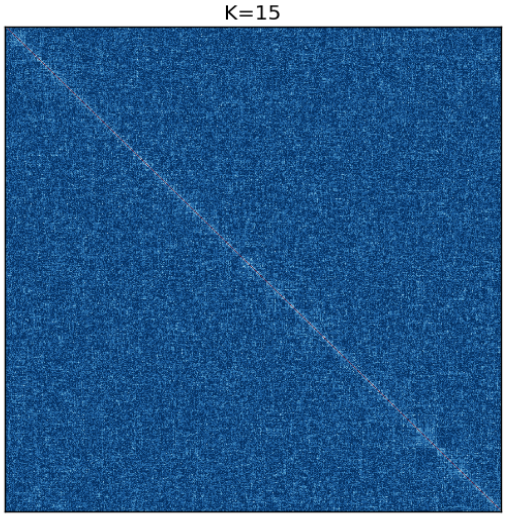
A ≈ W H











Next steps

- Change how we run logistical regression to get more repeatable results
- Continue to investigate ways to optimize our logistical regression accuracy via dimensionality reduction
- Investigate over-fitting problem: can we change algorithm parameters or implementation to avoid?