PCA by Hand

By William Gebhardt

Math from session 1

Transposes Matter

- When calculating eigenvalues they are the same for the transpose
- The eigenvectors change however
- Packages like numpy are flipped from conventional mathematics

$$A = \begin{bmatrix} -2 & -2 & 4 \\ -4 & 1 & 2 \\ 2 & 2 & 5 \end{bmatrix}$$

$$A^T = \begin{bmatrix} -2 & -4 & 2 \\ -2 & 1 & 2 \\ 4 & 2 & 5 \end{bmatrix}$$

The Same Eigenvalues

$$(\lambda - 3)(\lambda + 5)(\lambda - 6) = 0$$

$$A = \begin{bmatrix} -2 & -2 & 4 \\ -4 & 1 & 2 \\ 2 & 2 & 5 \end{bmatrix}$$

$$(\lambda - 3)(\lambda + 5)(\lambda - 6) = 0$$

$$\lambda = 3 \qquad \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix}$$

$$\lambda = -5 \qquad \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 2.8 \\ 2.2 \\ -1 \end{bmatrix}$$

$$\lambda = 6 \qquad \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0 \\ 1 \end{bmatrix}$$

Actually Doing PCA

Steps

1. Standardize data
    a. Zero-mean
    b. Standard deviation of 1
2. Compute the covariance matrix
3. Compute eigenvalue and vectors of covariance matrix
4. Order eigenvalues from largest to smallest
5. Compute desired variance captured
6. Reduce initial data set

Eigenvalue of a covariance matrix

$$[3.972, 1.702, 1.415, 1.073, 0.634, 0.564, 0.291, 0.22, 0.052, 0.076]$$

3.972	39.7%	0.564	5.6%
1.702	17%	0.291	2.9%
1.415	14.2%	0.22	2.2%
1.073	10.7%	0.052	0.5%
0.634	6.3%	0.076	0.8%

- Computed like normal
- Represent the variance of the data along their corresponding eigenvector
- The sum of all eigenvalues is the total variance across the data
- Proportions of the variance can be attributed to specific eigenvalues

3.972	39.7%	0.564	5.6%
1.702	17%	0.291	2.9%
1.415	14.2%	0.22	2.2%
1.073	10.7%	0.052	0.5%
0.634	6.3%	0.076	0.8%

Capturing Variance

By percent variance

- Select a threshold
- Add component starting with the most varied till passed

$$\text{threshold} = 80\%$$

$$\begin{aligned} 3.972 + 1.702 + 1.415 + 1.073 &= 8.16 \\ &= 81.6\% \end{aligned}$$

By number of components

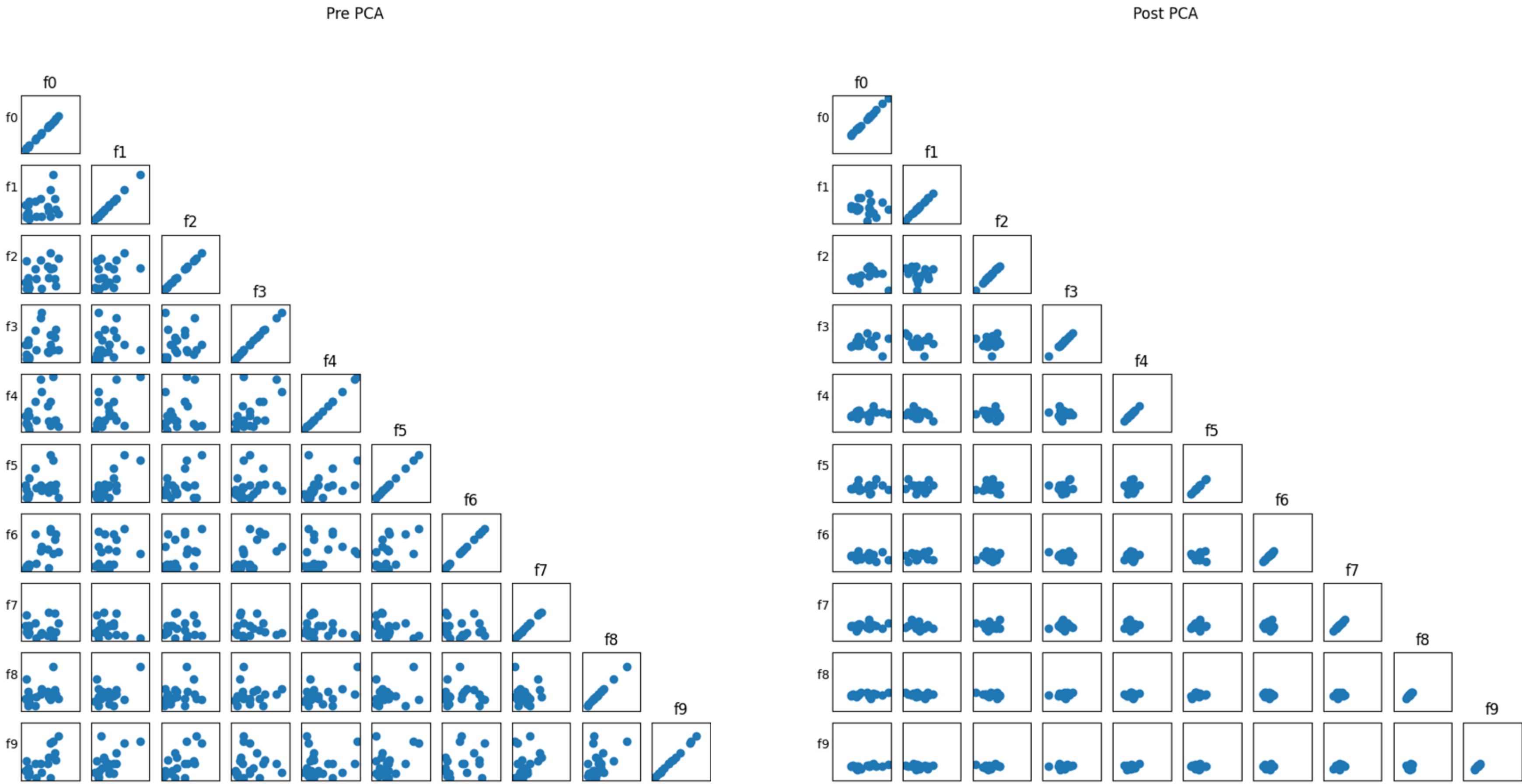
- Choose a number of components *n* to reduce the feature space too
- Add the largest *n* eigenvalues to get captured variance

$$n = 3$$

$$\begin{aligned} 3.972 + 1.702 + 1.415 &= 7.09 \\ &= 70.9\% \end{aligned}$$

Reducing the dataset

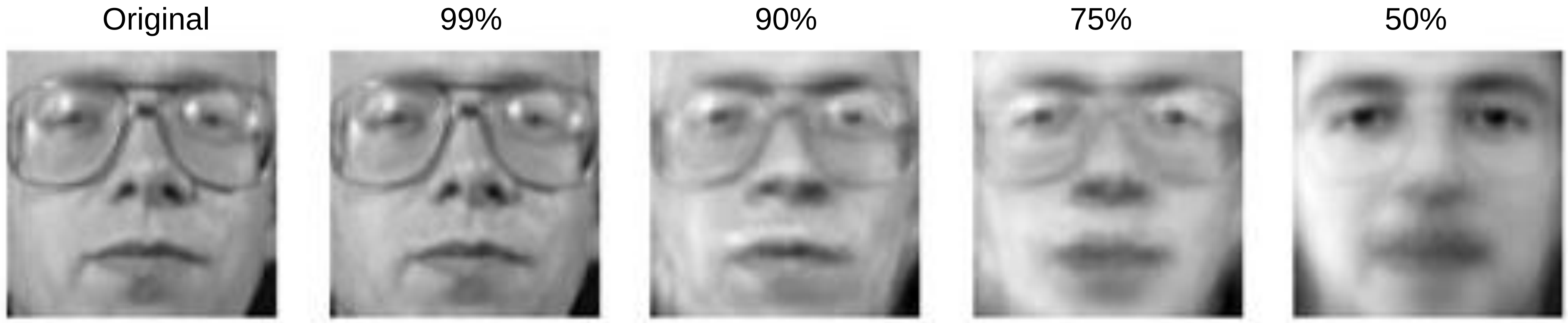
- Concatenate desired eigenvectors together
    - Forms (*num_features* x *num_components*)
- Take data and matrix multiply by the concatenated eigenvectors
    - (*num_points* x *num_features*)(*num_features* x *num_components*) = (*num_points* x *num_components)*
- Only the concatenated matrix of eigenvectors needs to be stores to use on future data


Pre PCA
f0
f1
f2
f3
f4
f5
f6
f7
f8
f9
Post PCA
f0
f1
f2
f3
f4
f5
f6
f7
f8
f9

Basic Code

Spot The Reproduction


Original
99%
90%
75%
50%