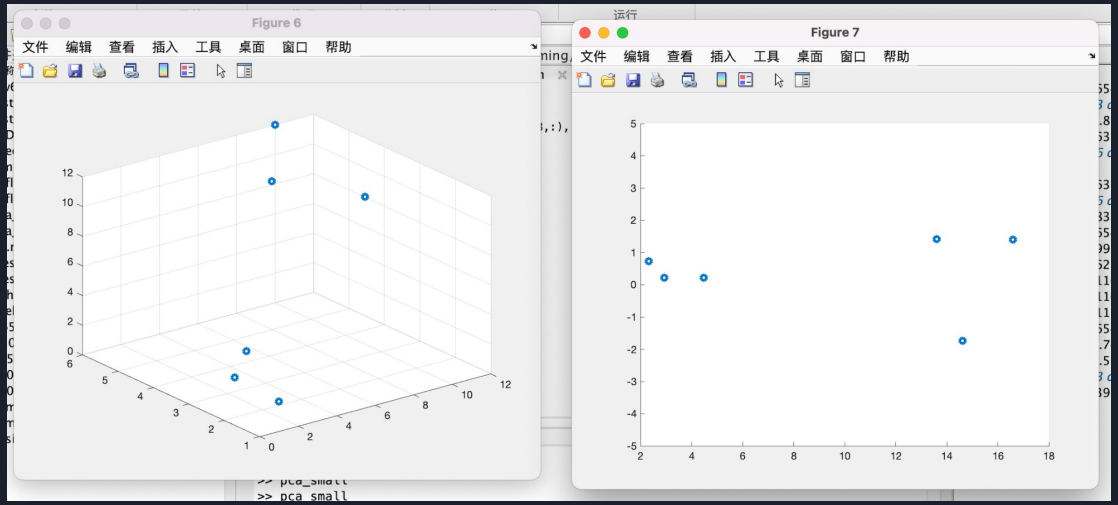
Principal Component Analysis

Ruihan Xu, Shiyu Liu, Zichen Gai, Zhanwang Zhou, Chenrui Hu

What is Principle Component Analysis (PCA)?

1. A dimensional reduction method for large data set
2. Transforming high dimensional data into a lower one but still keep the most information
3. Better to analyze, visualize the data set



A brief overview of PCA method with SVD

Methodology: Given matrix A n*m (n samples, m variable)

1. Standardization: get mean centered data matrix Z = (value - mean) / standard deviation
2. Covariance matrix: relationship between data, $C = B^T \times B$
3. Principal Component (PC): Eigenvalue and Eigenvector, **D & V**
4. Feature Vectors: vectors that contain high variance of data **W**
5. Reconstruction: plot the original data along PCs $Final\ Data = w^T \times A^T$

Importance of SVD: The key to reduce the dimension

1. Singular values are the importance of PCs
2. Eigenvectors V are the PCs

A brief overview of PCA method with SVD

Methodology: Given matrix A n*m (n samples, m variable)

1. Standardization: get mean centered data matrix Z = (value - mean) / standard deviation
2. Covariance matrix: relationship between data, $C = B^T \times B$

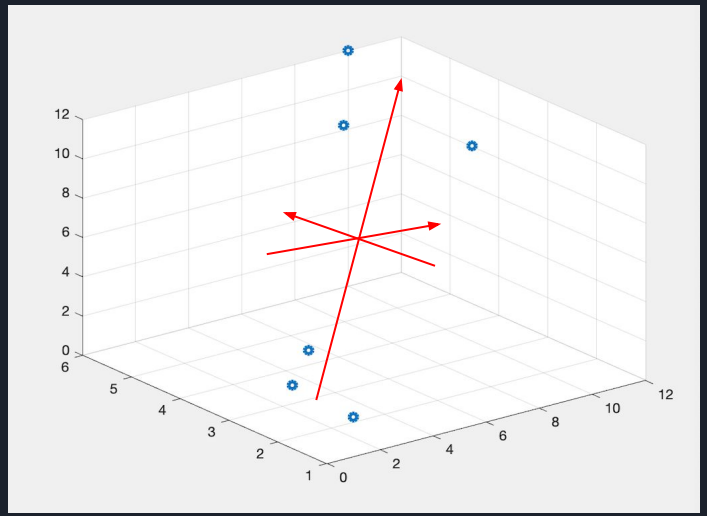
$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

Covariance Matrix for 3-Dimensional Data

A brief overview of PCA method with SVD

Methodology: Given matrix A n*m (n samples, m variable)

1. Standardization: get mean centered data matrix Z = (value - mean) / standard deviation
2. Covariance matrix: relationship between data, $C = B^T \times B$
3. Principal Component (PC): Eigenvalue and Eigenvector, **D & V**



A brief overview of PCA method with SVD

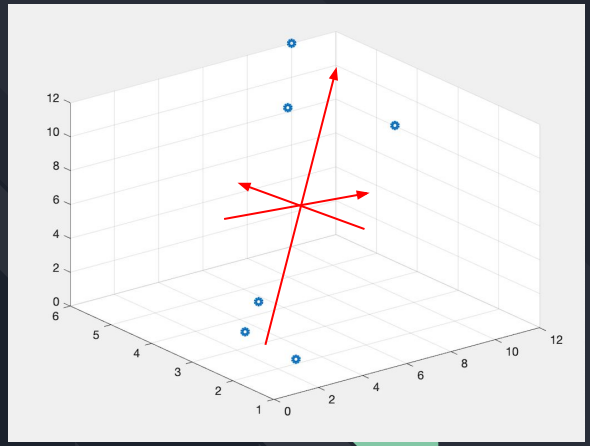
Methodology: Given matrix A n*m (n samples, m variable)

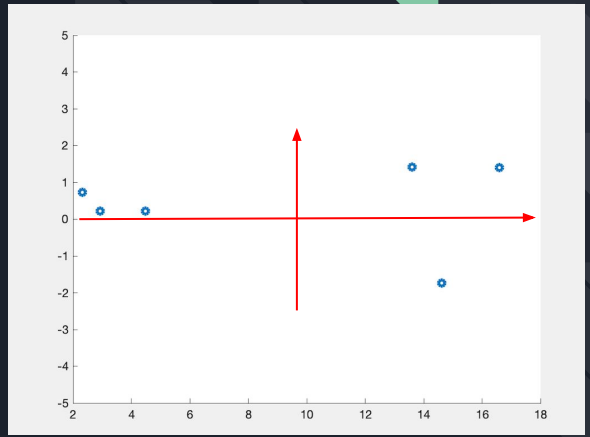
1. Standardization: get mean centered data matrix Z = (value - mean) / standard deviation
2. Covariance matrix: relationship between data, $C = B^T \times B$
3. Principal Component (PC): Eigenvalue and Eigenvector, **D & V**
4. Feature Vectors: vectors that contain high variance of data **W**
5. Reconstruction: plot the original data along PCs $Final\ Data = w^T \times A^T$

Importance of SVD: The key to reduce the dimension

1. Singular values are the importance of PCs
2. Eigenvectors V are the PCs

```m
data = [10, 11, 8, 3, 2, 1;
    6, 4, 5, 3, 2.8, 1;
    12,9,10,2.5,1.3,2];
X = data';
figure
scatter3(data(1,:),data(2,:),data(3,:),'LineWidth',3);

mean_matrix = mean(X);

B = X - ones(1,6)' * mean_matrix;
C = B' * B; % covariance matrix
[V, D] = eig(C);
lambda = eig(C);
S = sqrt(D);

v1 = V(:,3);
v2 = V(:,2);
newV = [V(:,3), V(:,2)];

figure
figure
for i = 1:size(X,1)
    x = newV(:,1)'*X';
    y = newV(:,2)'*X';
    scatter(x,0,'LineWidth',3);
    % scatter(x,y,'LineWidth',3);

end
ylim([-5,5]);
```



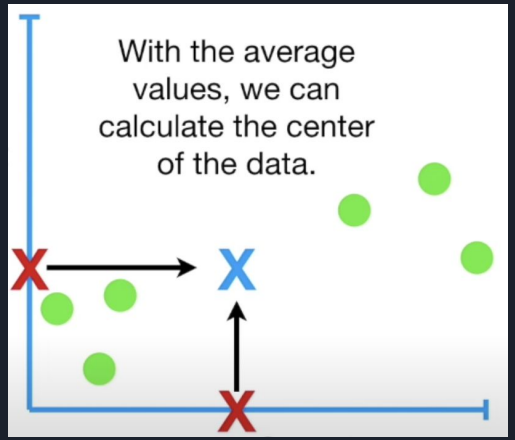


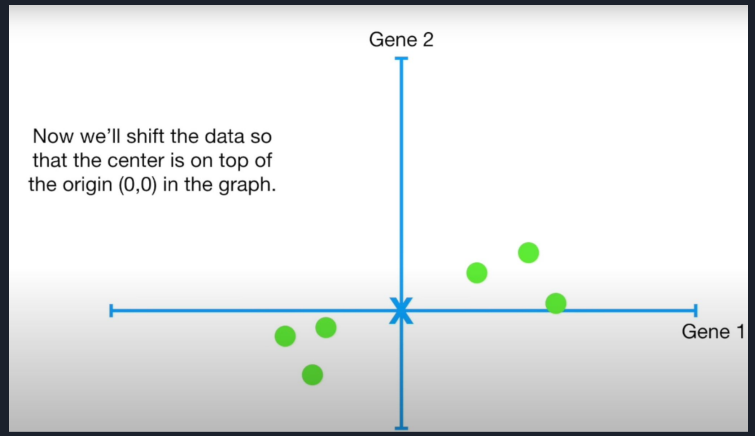
How To Compute? (Mathematics in PCA)

Let's start from the dataset that has only 2 variables (i.e in IR^2)

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
Gene 1	10	11	8	4	2	1
Gene 2	6	4	5	3	2.6	1

Step 1. Center the Data on the Plot





Z = (value - mean) / standard deviation

Step 2. Get the Covariance Matrix

After Centering Data, we can get the centered data matrix

B =

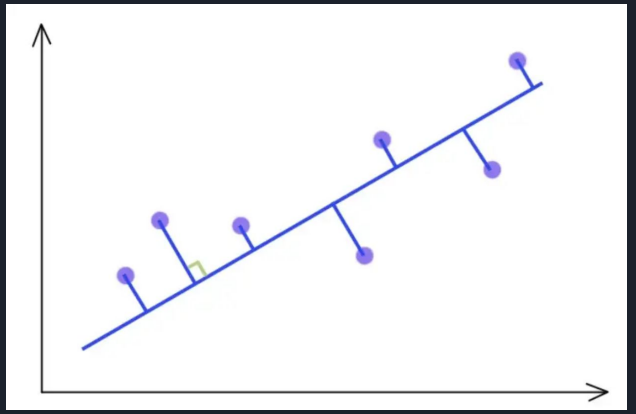
Col 1: Std variable Gene 1	Col 2: Std variable Gene 2
0.9428	1.3416
1.1785	0.2236
0.4714	0.7826
-0.4714	-0.3354
-0.9428	-0.5590
-1.1785	-1.4534

$$C = \begin{pmatrix} \sigma(x,x) & \sigma(x,y) \\ \sigma(y,x) & \sigma(y,y) \end{pmatrix}$$

Covariance Matrix C = (Transpose of B) *B =

5.0	4.3
4.3	5.0

Step 3. Principal Component (PC): Eigenvalue and Eigenvector, V, D



The Eigenvalues and Eigenvectors of the Covariance matrix C: (With Descending Order)

$\Lambda 1 = 9.3$ with eigenvector (1, 1) $\sigma 1 = 3.050$

$\Lambda 2 = 0.7$ with eigenvector (-1, 1) $\sigma 2 = 0.837$

If we rank the eigenvalues in descending order, we get $\lambda 1 > \lambda 2$, which means that the eigenvector that corresponds to the first principal component (PC1) is $v1$ and the one that corresponds to the second component (PC2) is $v2$.

$$PC1\% = \frac{3.050}{3.050 + 0.837} = 78.4\% \quad PC2\% = \frac{0.7}{9.3 + 0.7} = 21.6\%$$

After having the principal component, we compute the percentage of variance accounted for each component. We divide each singular value by the sum of singular value.
In this example, we find that PC1 and PC2 carry 78.4% and 21.6% of the variance of data.

Step 4. Feature Vector

From the previous step, we could find the percentage of data that each principal components represent.

We could discard PC2, which is the one of lesser significance and use PC1 to construct the data in 1 dimension:

[v1]

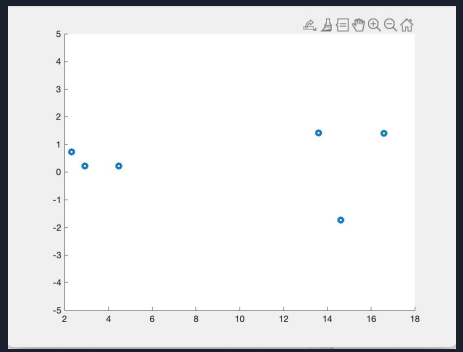


Image 1

Step 5. Reconstruction

For the last step, we use the feature vector to reorient the data from the original axes to the ones represented by the principal components. This can be done by multiplying the transpose of the original data set by the transpose of the feature vector

$$FinalDataSet = FeatureVector^T * StandardizedOriginalDataSet^T$$

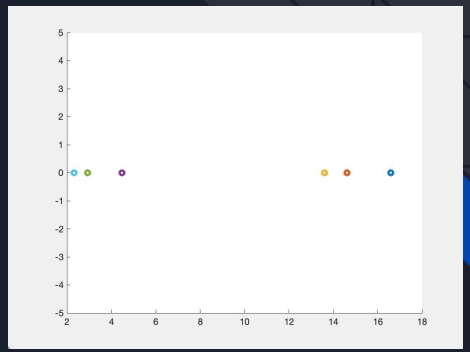


image 2

Example in IR^3

We will quickly go over the procedure of PCA in 3D, similar to the previous example but with higher dimension

	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6
Gene1	10	11	8	3	2	1
Gene2	6	4	5	3	2.8	1
Gene3	12	9	10	2.5	1.3	2

PCA in R^3

Step1: Center the data on the plot

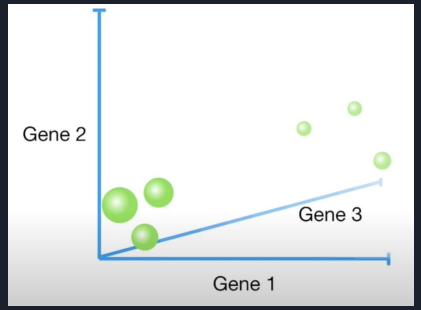


Image1

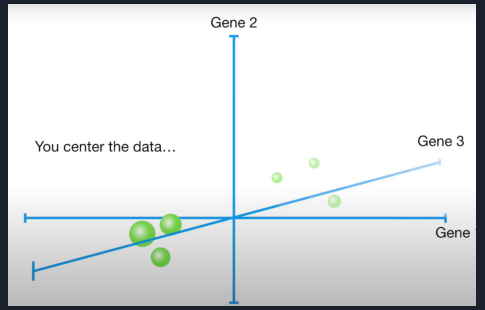


Image2

The centered data matrix becomes...
B = A-mean(A)/ standard deviation

B =

$$\begin{pmatrix} 4 & 2.4 & 5.87 \\ 5 & 0.4 & 2.87 \\ 2 & 1.4 & 3.87 \\ -2 & -0.6 & -3.63 \\ -4 & -1 & -4.83 \\ -5 & -2.6 & -4.13 \end{pmatrix}$$

Step 2. Find the Covariance Matrix

Covariance matrix equals the transpose of the centered matrix times centered matrix

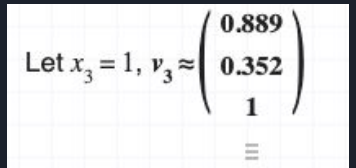
Covariance Matrix C = B'B =

$$\begin{pmatrix} 90 & 32.6 & 92.8 \\ 32.6 & 16 & 38.4 \\ 92.8 & 38.4 & 111.2334 \end{pmatrix}$$

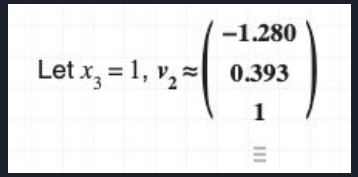
Step 3. Find Principal Component(PC)

We calculate the three eigenvalues and the corresponding eigenvectors, in which the eigenvectors associated with the largest eigenvalue is the PC1, that of second largest eigenvalue is the PC2, and so on…

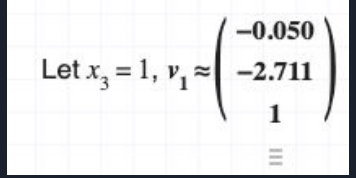
σ1 = 14.397



σ2 = 2.741



σ3 = 1.563



We could calculate the variation for each eigenvalue using singular value(sqrt of eigenvalue) divided by the sum of singular values.
PC1, PC2, and PC3 carry 76.97%, 14.66%, and 8.36% respectively of the variation of data

Step 4&5. Feature Vector & reconstruction

In the previous step, we found the PCs with their significance. That means that a 2-D graph, using just PC1 and PC2, would be a good approximation since it would account for around 91.5% of the variation in the data. Or a 1-D graph, using PC1 would also be a good approximation.

Take converting the 3d graph to a 2D graph as an example.

What we need to do is to :

1. Eliminate everything on the graph except data, PC1, and PC2.
2. Project points onto PC1.
3. Project points onto PC2.
4. Rotate PC1 to horizontal and PC2 to vertical(easier to look at) and recover the points in our new PCA plot.

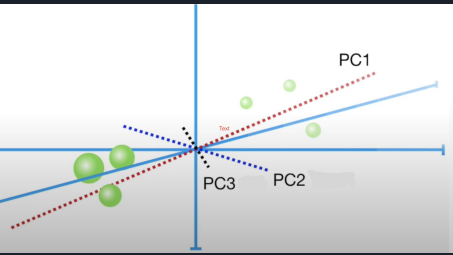


Image1

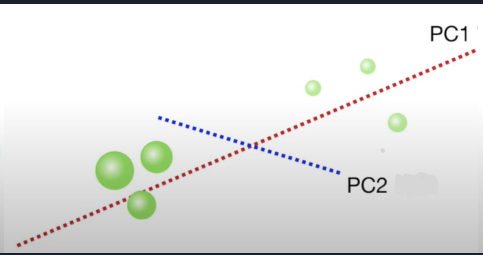


image2

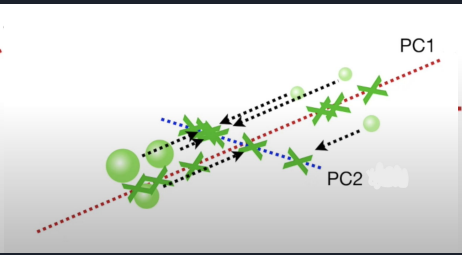


image3

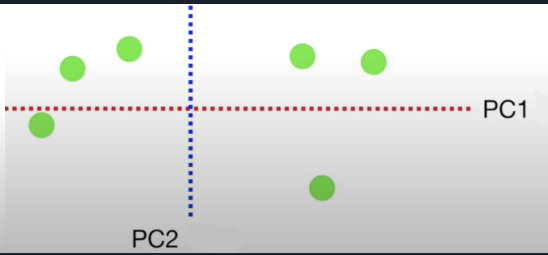


image4

Real World Application 1: Visualizing Country clustering

Country_informtion dataset:

- 129 countries
- 12 sociological variables

Selected two principal components with largest EVR for visualization in 2D

Explained variance ratio:

```
[0.52152711 0.13897869 0.10503116 0.0691413  0.04954995 0.03272344]
```

Singular values:

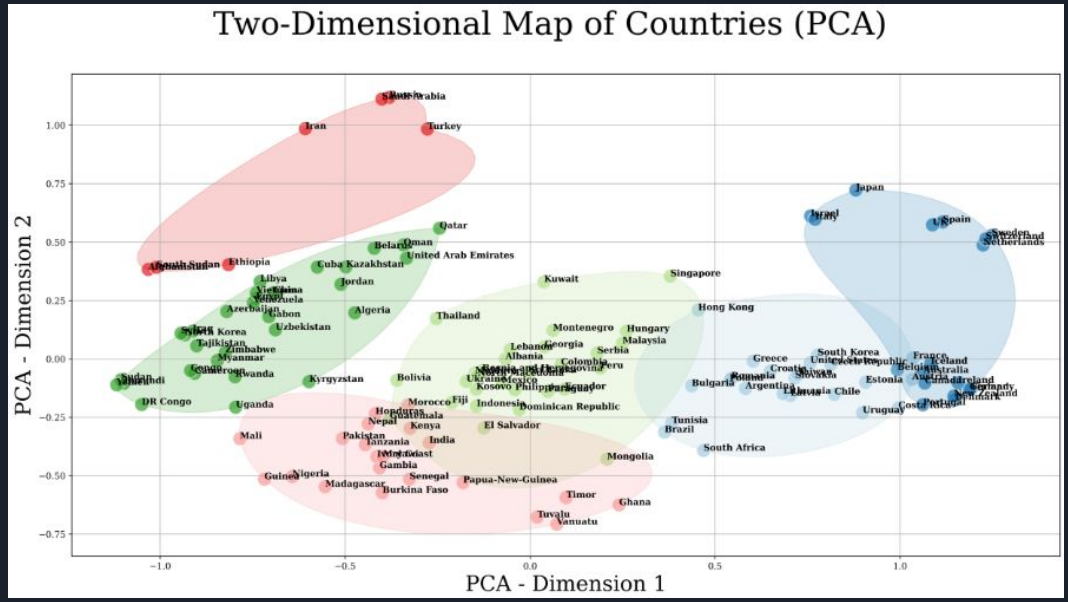
```
[7.75630969 4.00396981 3.48077352 2.82413515 2.39077275 1.94288093]
```

Variable weights:

```
      gini_index  corruption_perceptions_index  freedom_house       hdi  \
PC-1   -0.060241                      0.316372       0.532343  0.389982
PC-2   -0.097666                      0.127682      -0.375216  0.467918

      press_freedom  democracy_economist  populism  \
PC-1       0.343360             0.459797  0.131003
PC-2      -0.204977            -0.233690 -0.013420

      effective_coverage_of_health_services_index  trust_in_news_media  \
PC-1                                     0.288932             0.027475
PC-2                                     0.311854            -0.096228

      trust_in_government  trust_in_science  colonized
PC-1            -0.140257          0.080304  -0.063059
PC-2            -0.139149          0.053185  -0.623629
```

Visualization

- 12 to 2 dimensions
- PCs representing 2 axis
- Center around origin
- 6 clusters
- Countries of the same group tend to have common traits
- Synthesize original variables, reduce complexity



Real World Application 2: Face Recognition

Step 1: convert the picture set into a matrix

```
X = np.genfromtxt("faces.txt", delimiter=None)
plt.figure()
print(X.shape)
```

```
(4916, 576)
```

Step 2: find out mean-subtracted matrix

```
mu = np.mean(X, axis=0, keepdims=True)
img = np.reshape(mu,(24,24))
plt.imshow( img.T , cmap="gray")
X0 = X - mu
```

Step 3: SVD

```
U, S, Vh = np.linalg.svd(X0, full_matrices=False)
Sigma = np.diag(S)
```

Step 4: extract top K eigenface pictures

```
W = U.dot(Sigma)
imgs = []
for j in [1, 2, 3]:
    alpha = 2 * np.median(np.abs(W[:,j]))
    img = np.reshape(mu+alpha * Vh[j,:],(24,24))
    plt.imshow( img.T , cmap="gray")
    plt.show()
```

Step 5: project the given picture onto eigenface

```
print(Vh[:3,:].dot(X0[1,:]))
```

```
[-223.93674278  431.08227752  919.10438615]
```

Thank you!

Work cited

Principal Component Analysis (PCA) Explained | Built In. https://builtin.com/data-science/step-step-explanation-principal-component-analysis. Accessed 6 Dec. 2022.

Principal Component Analysis (PCA) [Matlab]. Directed by Steve Brunton, 2020. *YouTube*, https://www.youtube.com/watch?v=VqjJ5YYt78Y.

StatQuest: Principal Component Analysis (PCA), Step-by-Step. Directed by StatQuest with Josh Starmer, 2018. *YouTube*, https://www.youtube.com/watch?v=FgakZw6K1QQ.