PCA_report

Weijin Zhu Kun Lin

October 2020

1 Principal Component Analysis (PCA)

1.0.1 Goals

The main goal of performing PCA algorithm is to:

- Transfer correlated dimensions into new set of uncorrelated dimensions
- Map the original data into a lower dimension space

1.1 Implementations

1. Step 1: Calculate the mean and compute the mean vector $\bar{X}$.Then adjust the original data by the mean to center the original data around the mean $x' = x - \bar{X}$
2. Step 2: Calculate the covariance matrix of the adjusted data points using the formula:
$$S = \frac{1}{n-1}\mathbf{x'}^{\intercal}\mathbf{x'}$$
3. Step 3: Find the eigenvectors and eigenvalus of S based on the formula:
$$Sa = \lambda a$$
and select top eigenvalues and its corresponding eigenvectors depends on the dimensionality we want it to reduce to.
4. Step 4: Transform the original data into a new dimension using the equation below
$$y_j = Xa_j$$
where a_j is the eigenvectors

1.2 Procedures

Given a file of data, each column represents the feature of that disease given on the last column, use *pca_a.txt* as the example:

Example				
X_1	X_2	X_3	X_4	Disease
2.60	5.40	1.90	6.40	Asthma
3.30	6.00	2.20	6.70	Arrhythmia
3.80	2.00	0.80	5.40	Hypertension
...	...	...	...	...
3.60	1.70	0.60	5.1	Hypertension

1. Step 1: Calculate the mean for every column (X_1, X_2, X_3, X_4) to compute the mean vector $\bar{X}$.Then adjust the original data by the mean to center the original data around the mean $x' = x - \bar{X}$

$$\bar{X} = [3.454, 4.1587, 1.5987, 6.243]$$

Example				
X'$_1$	X'$_2$	X'$_3$	X'$_4$	Disease
-0.854	1.24	0.301	0.157	Asthma
-0.154	1.84	0.6013	0.457	Arrhythmia
0.346	-2.1587	-0.7987	-0.843	Hypertension
...	...	...	...	...
0.146	-2.46	-0.9987	-1.143	Hypertension

2. Step 2: Calculate the covariance matrix of the adjusted data points using the formula:

$$S = \frac{1}{n-1}\mathbf{x'}^{\intercal}\mathbf{x'}$$

$$\mathrm{S} = \begin{bmatrix} 0.18800403 & -0.32171275 & -0.11798121 & -0.03926846 \\ -0.32171275 & 3.11317942 & 1.29638747 & 1.27368233 \\ -0.11798121 & 1.29638747 & 0.58241432 & 0.5169038 \\ -0.03926846 & 1.27368233 & 0.5169038 & 0.68569351 \end{bmatrix}$$

3. Step 3: Find the eigenvectors and eigenvalus of S. Choose top 2 largest eigenvalues with its corresponding eigenvectors.

$$\text{Eigenvalues} = \begin{bmatrix} 4.22484077 & 0.24224357 & 0.07852391 & 0.02368303 \end{bmatrix}$$

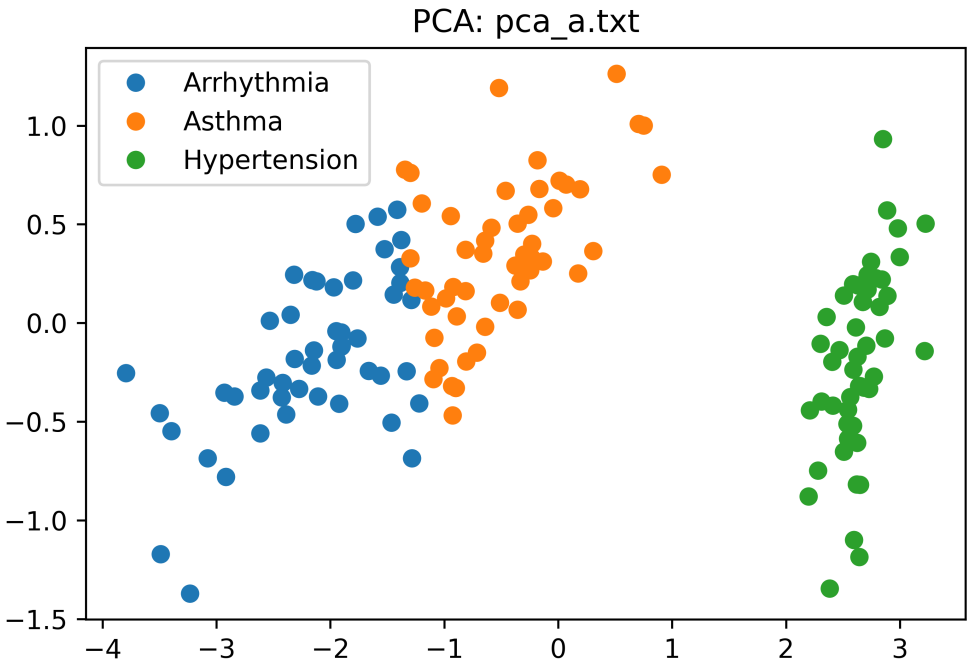
$$\text{Eigenvectors} = \begin{bmatrix} 0.08226889 & -0.72971237 & -0.59641809 & 0.32409435 \\ -0.85657211 & 0.1757674 & -0.07252408 & 0.47971899 \\ -0.35884393 & 0.07470647 & -0.54906091 & -0.75112056 \\ -0.36158968 & -0.65653988 & 0.58099728 & -0.31725455 \end{bmatrix}$$

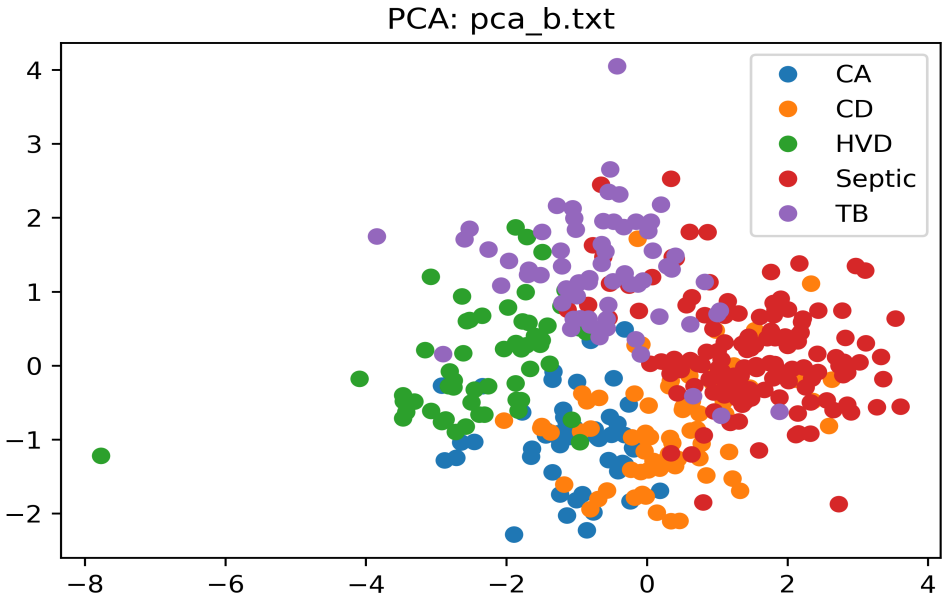
Step 4: Transform the original data into a new dimension

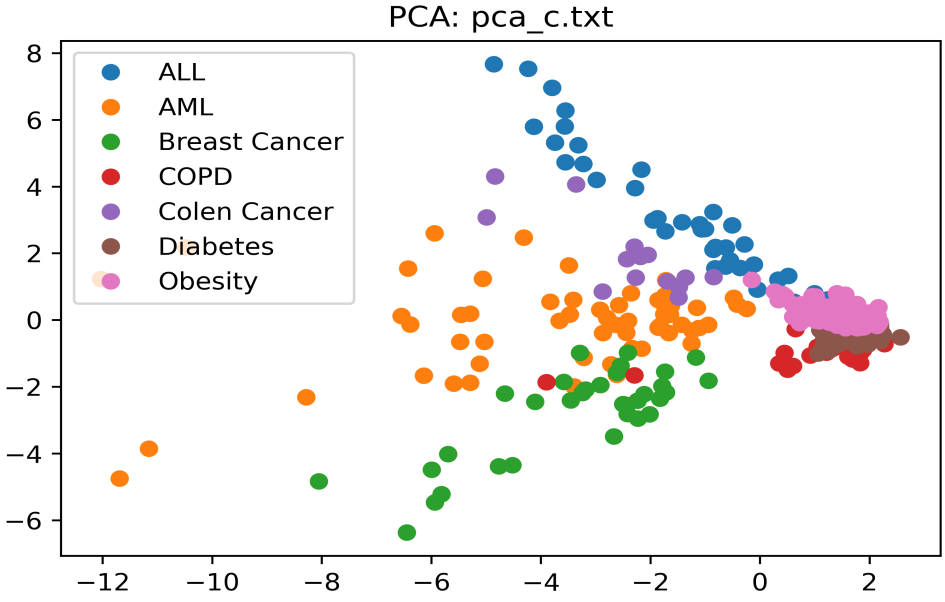
$$\begin{bmatrix} -1.298330 & 0.761014 \\ -1.970815 & 0.181126 \\ 2.469056 & -0.137887 \\ -2.840961 & -0.372743 \\ 2.998296 & 0.334308 \\ \dots & \dots \\ 2.889820 & 0.137346 \end{bmatrix}$$

2 Results

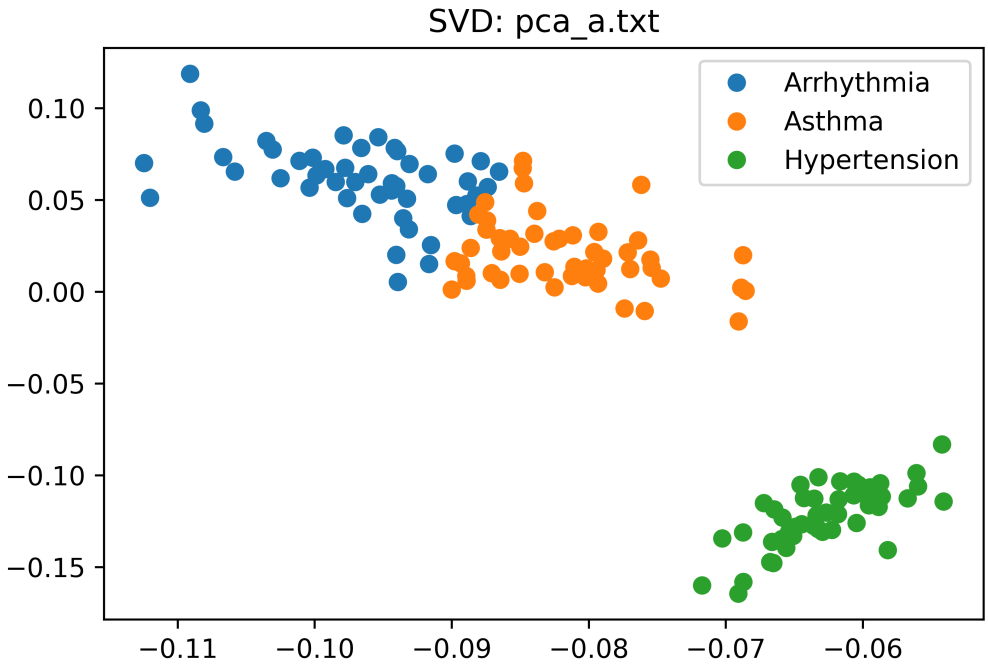
2.1 PCA

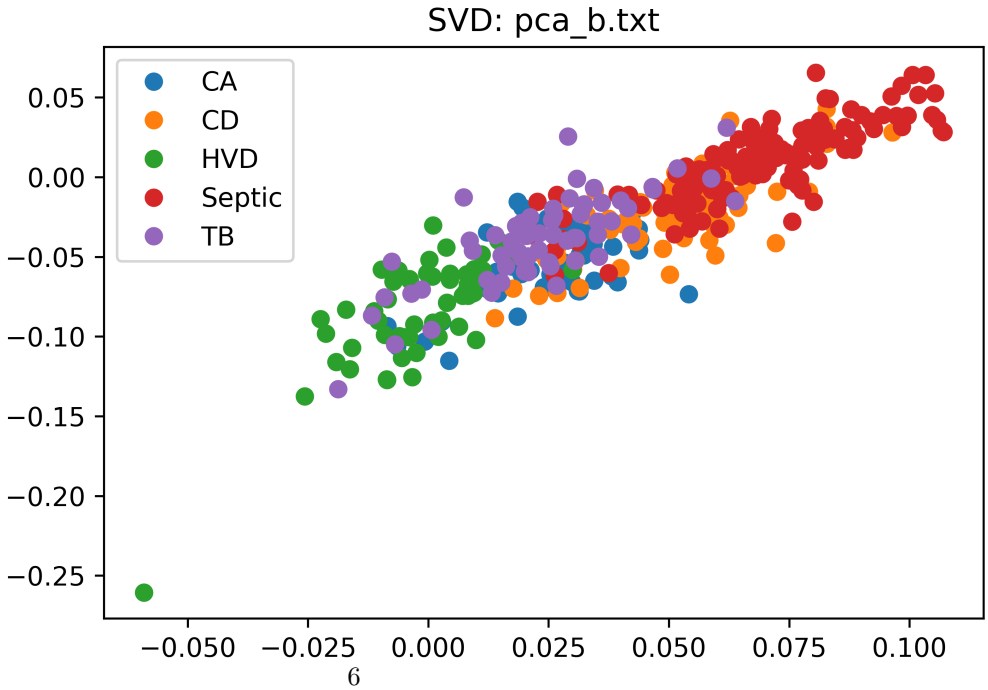


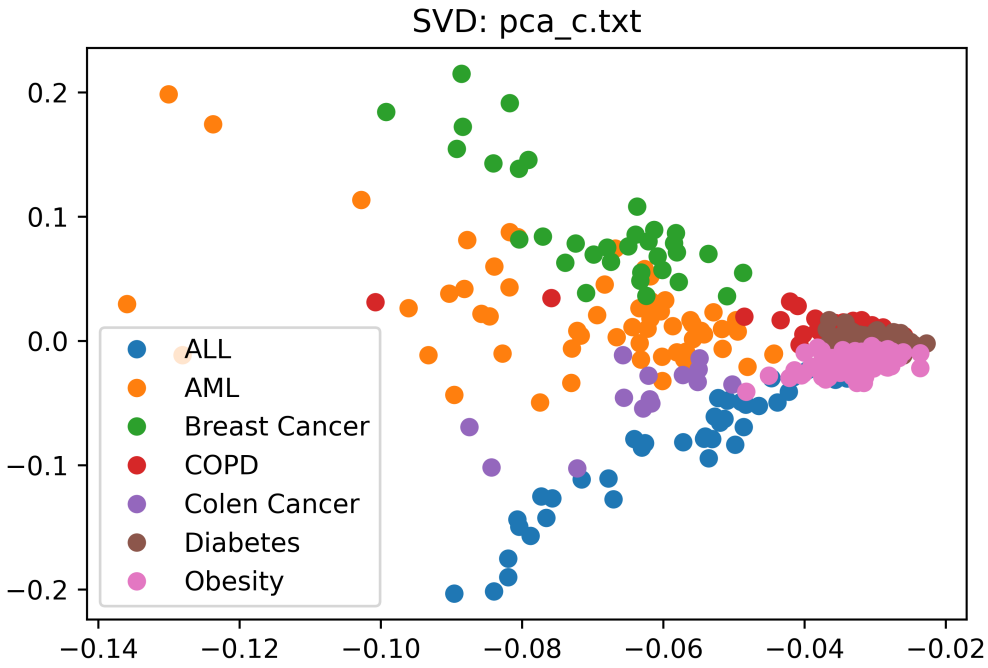

PCA: pca_b.txt
CA
CD
HVD
Septic
TB
4
3
2
1
0
−1
−2
−8
−6
−4
−2
0
2
4


PCA: pca_c.txt
ALL
AML
Breast Cancer
COPD
Colen Cancer
Diabetes
Obesity
8
6
4
2
0
−2
−4
−6
−12
−10
−8
−6
−4
−2
0
2

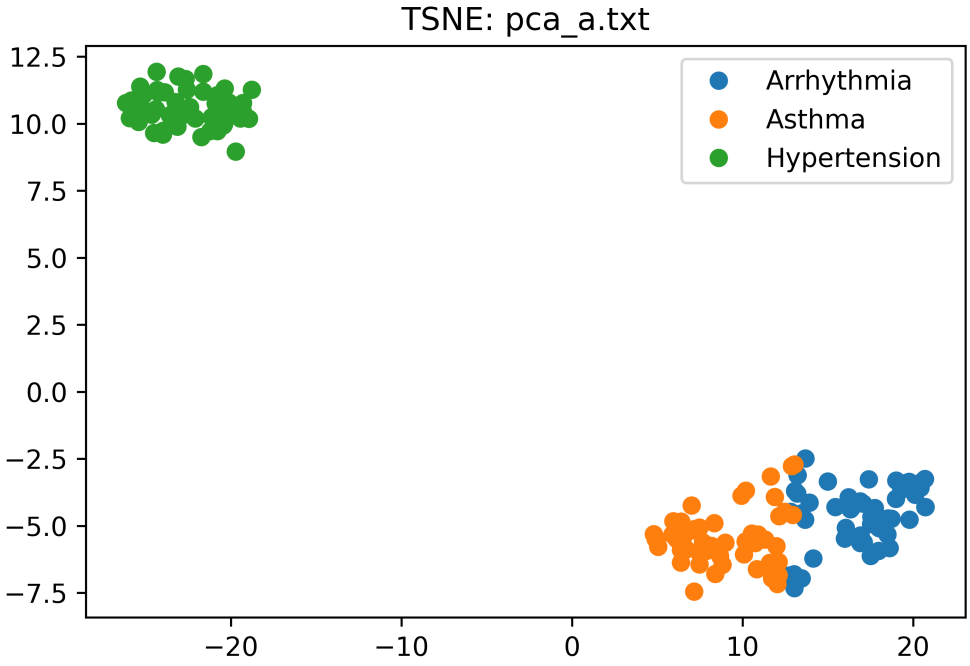
2.2 SVD

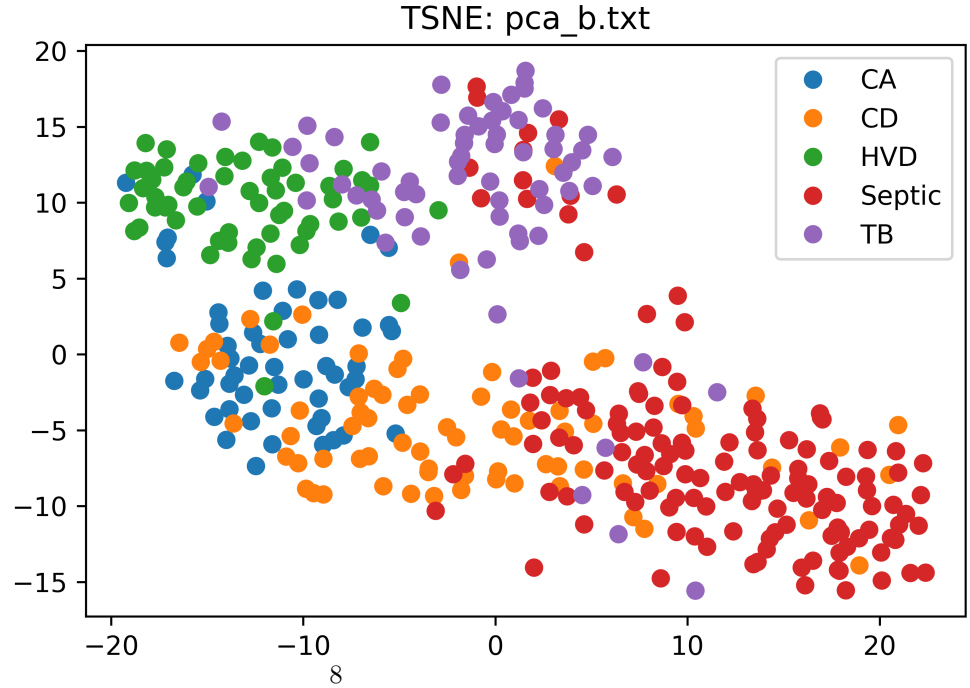

SVD: pca_a.txt
Arrhythmia
Asthma
Hypertension
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.11
−0.10
−0.09
−0.08
−0.07
−0.06

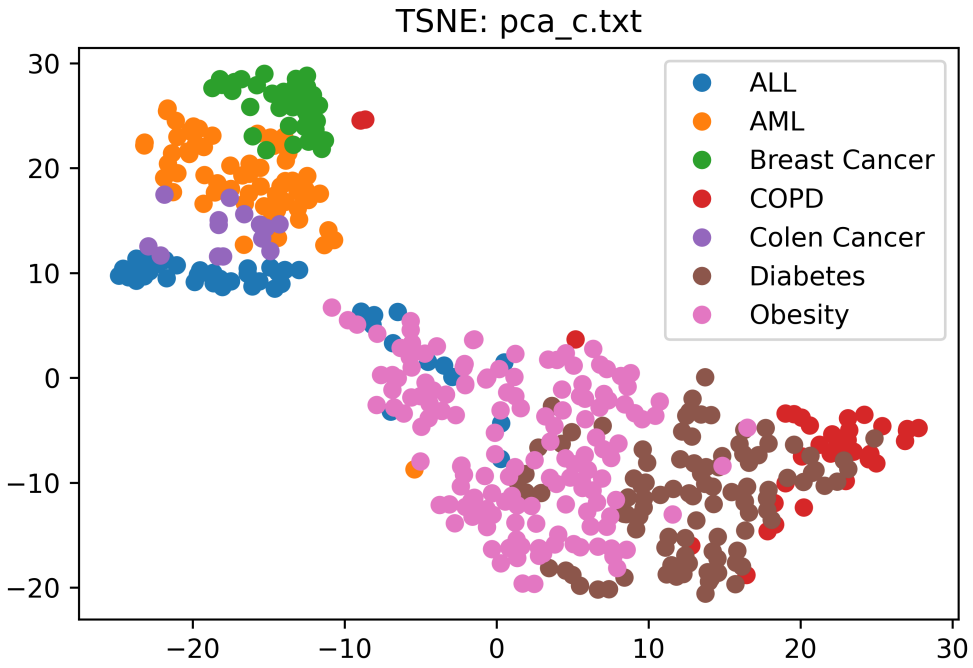

SVD: pca_b.txt
CA
CD
HVD
Septic
TB
0.05
0.00
−0.05
−0.10
−0.15
−0.20
−0.25
−0.050
−0.025
0.000
0.025
0.050
0.075
0.100


SVD: pca_c.txt
0.2
0.1
0.0
−0.1
−0.2
−0.14
−0.12
−0.10
−0.08
−0.06
−0.04
−0.02
ALL
AML
Breast Cancer
COPD
Colen Cancer
Diabetes
Obesity

2.3 TSNE


TSNE: pca_a.txt
Arrhythmia
Asthma
Hypertension
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−20
−10
0
10
20


TSNE: pca_b.txt
CA
CD
HVD
Septic
TB
20
15
10
5
0
−5
−10
−15
−20
−10
0
10
20



3 Observations

In order to calculate the principal components for PCA, we need the result of SVD. That is, by observing the results from these two algorithms, we discovered that if the existing SVD package takes in the data frame as the input, the result will be different from PCA algorithm. However, like what we did in PCA if we adjust the original data by the mean, we center the original data around the mean to be our new data frame. When SVD takes this new matrix as its input, the result is the mirror of PCA. The result of TSNE is different every time, the reason is that TSNE is trying to minimize the KL divergence between the joint probabilities of low-dimensional and high dimensional data, so it is a probabilistic approach.