SRECHARAN SSELVAM

Homework 12

Instructions

This homework contains **5** concepts and **3** programming questions. In MS word or a similar text editor, write down the problem number and your answer for each problem. Combine all answers for concept questions in a single PDF file. Export/print the Jupyter notebook as a PDF file including the code you implemented and the outputs of the program. Make sure all plots and outputs are visible in the PDF.

Combine all answers into a single PDF named andrewID_hw12.pdf and submit it to Gradescope before the due date. Refer to the syllabus for late homework policy. Please assign each question a page by using the "Assign Questions and Pages" feature in Gradescope.

Here is a breakdown of the points for programming questions:

Name	Points
M12-L1-P1	15
M12-L2-P1	25
M12-HW1	50

Problem 1 (2 points)

What would the dimension of the covariance matrix be for the following data:

(Choose one)

1. 2 x 2
2. 6 x 6
3. 12 x 12
4. 20 x 20

x_1	x_2	x_3	x_4	x_5	x_6
-7.55	5.85	11.88	1.99	6.39	3.05
-10.93	6.56	8.96	-0.89	7.43	4.07
-9.44	6.37	9.86	-0.62	7.73	2.88
-1.83	0.53	8.55	-6.21	-8.05	5.13
6.38	0.47	-6.72	2.71	-5.24	-2.11
7.85	-0.17	-8.48	1.40	-7.62	-3.71
9.17	0.70	-7.45	2.09	-6.13	-4.66
0.76	1.97	8.46	-5.47	-7.57	3.33
-11.58	6.13	9.34	0.21	9.00	3.03
-8.41	5.29	10.13	-0.97	7.48	5.11
-7.87	5.48	10.50	1.71	6.04	3.79
-0.84	0.23	7.99	-6.91	-7.59	3.11
1.06	-0.56	7.47	-7.12	-6.31	3.82
7.43	1.26	-8.13	1.30	-5.78	-6.79
0.59	0.88	7.85	-6.20	-8.18	3.94
7.35	1.04	-5.98	1.61	-5.69	-5.54
1.01	1.40	9.87	-5.62	-7.74	4.08
8.47	2.80	-7.24	0.93	-5.39	-4.60
8.00	1.39	-6.57	0.53	-2.77	-7.12
-10.92	7.00	8.96	-1.30	6.90	4.82

Problem 2 (2 points)

Provided the following eigenvalues and eigenvectors e_1 and e_2, what are the values *i*, *j*, *k*, that comprise the unit normalized third eigenvector, e_3?
(Text entry for each i, j, k)

$\lambda_1 = 16$
$\lambda_2 = 4$
$\lambda_3 = 0$
$e_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \end{bmatrix}$
$e_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 \end{bmatrix}$
$e_3 = \begin{bmatrix} i & j & k \end{bmatrix}$

Problem 3 (2 points)

The eigenvalues of the covariance matrix from the data in the first concept question are included below. Which components should be used to explain at least 80% of the variance in the data?

$\lambda_1 = 160.30$

$\lambda_2 = 44.31$

$\lambda_3 = 1.86$

$\lambda_4 = 1.47$

$\lambda_5 = 0.62$

$\lambda_6 = 0.49$

Multiple choice (select all that apply)

- PC1
- PC2
- PC3
- PC4
- PC5
- PC6

Problem 4 (2 points)

What should the dimension of the covariance matrix be for the following data:
(Choose one)

1. 2 x 2
2. 6 x 6
3. 10x10

4. 20 x 20

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
-9.25	2.84	-9.38	0.66	5.71	-2.23	8.76	-5.37	-2.56	1.25
-10.24	3.23	-8.34	-0.70	5.53	-2.72	8.70	-4.77	-2.61	0.44
2.36	-10.36	5.22	-2.26	7.44	-4.88	-4.87	1.83	-8.76	-7.48
-7.84	5.72	-2.35	8.14	-6.54	10.40	-2.19	-2.51	-3.84	-1.19
-7.51	5.07	-2.21	6.73	-7.42	8.83	-4.00	-2.65	-3.57	-0.89
0.49	-8.68	4.84	0.05	6.40	-4.71	-4.96	2.05	-7.59	-6.18

Problem 5 (2 points)

Select the following statements about t-SNE which are true: (Multiple choice, select all that apply)

1. t-SNE can be used to project unseen high dimensional data into a reduced feature space
2. t-SNE preserves global structure and distances between data points by computing pairwise similarities
3. Like PCA, t-SNE is a linear dimensionality reduction technique that is used to reduce high dimensional data to a low dimensional feature space
4. t-SNE is a non-linear dimensionality technique that can learn embeddings of manifolds

ANSWERS :-

PROBLEM 1:-

OPTION: 2 6×6

PROBLEM 2:-

components, $i = 0$, $j = 0$ and $k = 1$

$[0, 0, 1]$

PROBLEM 3:-

PC1 $\rightarrow$ 76.7% variance

PC2 $\rightarrow$ 97.9% variance

PROBLEM 4:-

OPTION: 3 10×10

PROBLEM 5:-

option (2) & (4) are correct

M12-L1 Problem 1

This problem is intended to demonstrate PCA on a small 2D dataset. This will emphasize how PCs are computed and what they mean.

In [11]:
```
import numpy as np
import matplotlib.pyplot as plt

X = np.array([[2.5, 2.4],[0.5, 0.7],[2.2, 2.9],[1.9, 2.2],[3.1, 3. ],
              [2.3, 2.7],[2., 1.6],[1., 1.1],[1.5, 1.6],[1.1, 0.9]])
```

Computing the Principal Components

First, compute the principal components of the dataset by following these steps:

1. Compute `M` (1×2), the mean of each dimension in `X`
2. Compute `S` (2×2), the covariance matrix of `X` (see `np.cov`)
3. Report `w`, the 2 eigenvalues of `S` (see `np.linalg.eig`)
4. Get `e1` and `e2`, the eigenvectors corresponding to the elements of `w`

The principal components in this problem are then `e1` and `e2`.

In [12]:
```
print('X:\n', X)

# YOUR CODE GOES HERE: Compute M
M = np.mean(X,axis=0)
print('\nMean of each dimension:\n', M)

# YOUR CODE GOES HERE: Compute S
S = np.cov(X.T)
print('\nCovariance Matrix:\n', S)


# YOUR CODE GOES HERE: Compute w
```

```
w,v = np.linalg.eig(S)
print('\nEigenvalues of covariance matrix:\n',w)

# YOUR CODE GOES HERE: Compute e1, e2
sort_ind = np.argsort(w)[::-1]
w_sort = w[sort_ind]
v_sort = v[:,sort_ind]
e1 = v_sort[:, 0]
e2 = v_sort[:, 1]
print('\nPrincipal Components:')
print('e1:',e1)
print('e2:',e2)
```

```
X:
 [[2.5 2.4]
 [0.5 0.7]
 [2.2 2.9]
 [1.9 2.2]
 [3.1 3. ]
 [2.3 2.7]
 [2.  1.6]
 [1.  1.1]
 [1.5 1.6]
 [1.1 0.9]]

Mean of each dimension:
 [1.81 1.91]

Covariance Matrix:
 [[0.61655556 0.61544444]
 [0.61544444 0.71655556]]

Eigenvalues of covariance matrix:
 [0.0490834  1.28402771]

Principal Components:
e1: [-0.6778734  -0.73517866]
e2: [-0.73517866  0.6778734 ]
```

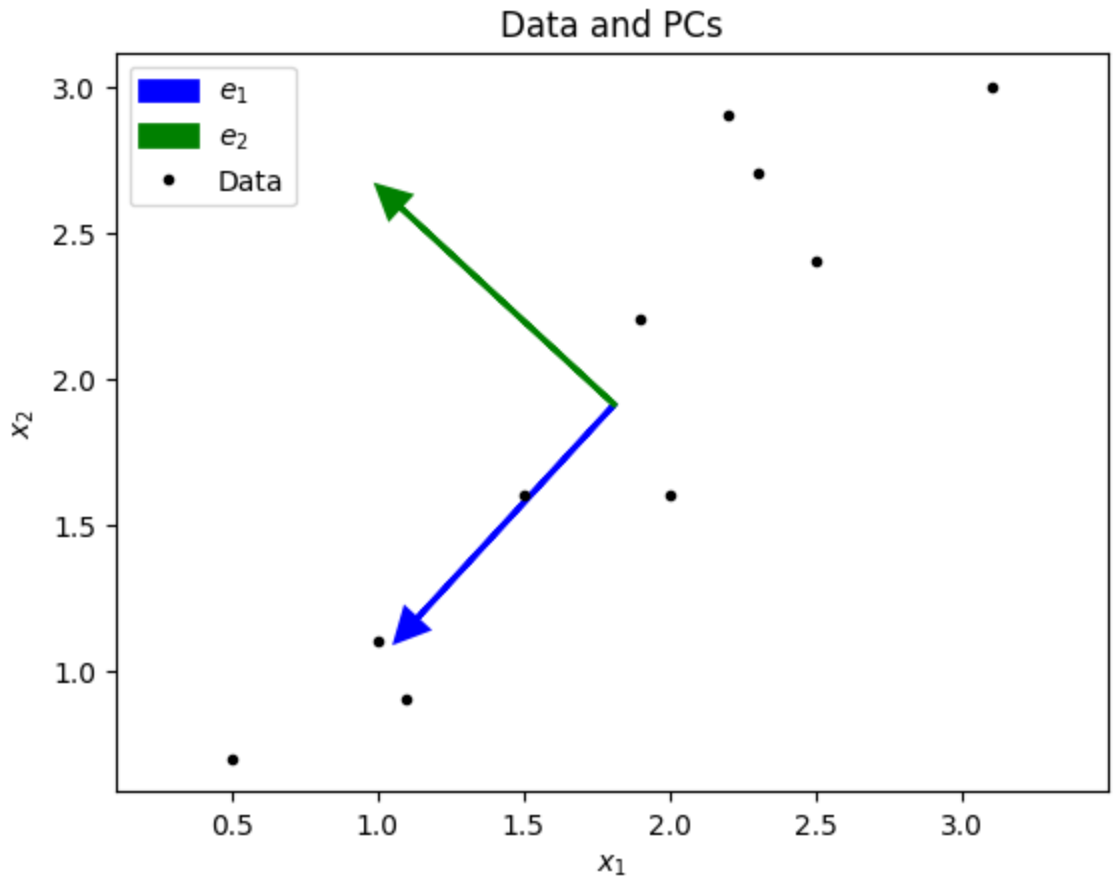
Plotting data with principal components

Complete the code below to plot the original data with principal components represented as unit vector arrows.

In [14]:
```
plt.figure()
plt.title("Data and PCs")

e1, e2 = e1.flatten(), e2.flatten()
plt.arrow(M[0],M[1],e1[0],e1[1], color="blue", linewidth=2, head_width=0.1, head_length=0.1, label="$e_1$")
plt.arrow(M[0],M[1],e2[0],e2[1], color="green", linewidth=2, head_width=0.1, head_length=0.1, label="$e_2$")
plt.plot(X[:,0],X[:,1],'.',color="black", label="Data")

plt.xlabel("$x_1$")
plt.ylabel("$x_2$")
plt.legend()
plt.axis("equal")
plt.show()
```



Plotting transformed data

Now, transform the data with the formula $a_i = (x - \mu) \bullet e_i$.

Print the transformed data matrix columns `a1` and `a2`.

Then plot the transformed data on $e_1 - e_2$ axes.

In [18]:
```
# YOUR CODE GOES HERE: Compute a1, a2
a1 = (X-M).dot(e1)
a2 = (X-M).dot(e2)
```

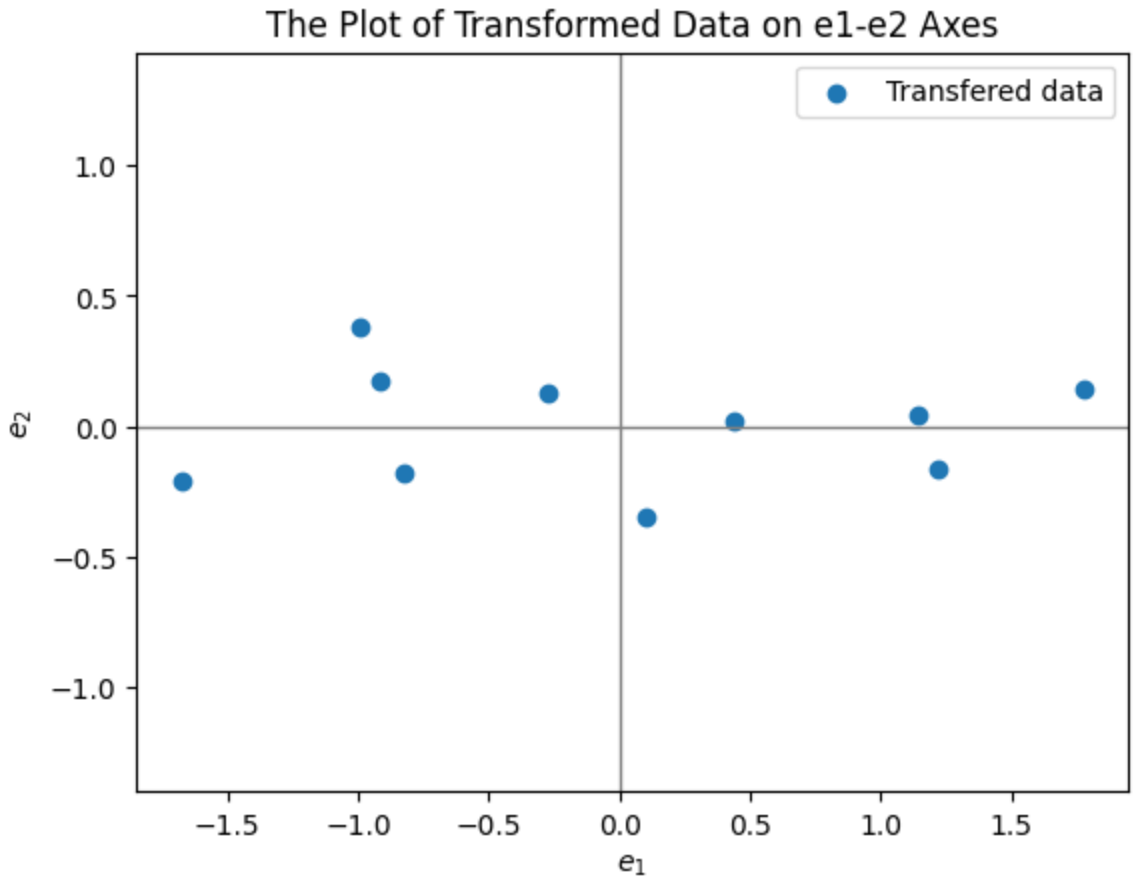
```
print("a_1 = ",a1)
print("a_2 = ",a2)

plt.figure()
plt.title("Transformed data")

e1, e2 = e1.flatten(), e2.flatten()
# YOUR CODE GOES HERE: Plot transformed data
plt.scatter(a1, a2, label ="Transfered data")
plt.title('The Plot of Transformed Data on e1-e2 Axes')
plt.axhline(0, color = 'grey', lw=1)
plt.axvline(0, color = 'grey', lw=1)
plt.legend()

plt.xlabel("$e_1$")
plt.ylabel("$e_2$")
plt.axis("equal")
plt.show()
```

```
a_1 =  [-0.82797019  1.77758033 -0.99219749 -0.27421042 -1.67580142 -0.9129491
  0.09910944  1.14457216  0.43804614  1.22382056]
a_2 =  [-0.17511531  0.14285723  0.38437499  0.13041721 -0.20949846  0.17528244
 -0.3498247   0.04641726  0.01776463 -0.16267529]
```


The Plot of Transformed Data on e1-e2 Axes
Transfered data
1.0
0.5
0.0
−0.5
−1.0
e_2
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
e_1

M12-L1 Problem 2

Sometimes the dimensionality is greater than the number of samples. For example,in this problem X has 19 features, but there are only 4 data points. You will need to use the alternate PCA formulation in this case. Follow the steps in the cells below to implement this method.

In [50]:
```
import numpy as np
import matplotlib.pyplot as plt

X = np.array([  [-2,  1,  2, -3,  4,  1,  0,  3,  0,  2,  1,  1,  2,  3, -2, -3, 2,  1,  0],
                [ 1,  2, -4,  2, -4,  2,  5,  2,  2,  1, -3,  0,  0,  1, -2,  1, 1, -3, -2],
                [ 1, -3,  2,  1,  0, -3, -5, -1,  3,  3, -2, -3, -2, -1,  1,  0, 5,  4,  2],
                [ 3, -1,  0,  2,  2, -5, -4, -1,  2, -1,  3,  4,  4,  2,  1,  2, -2,  1, -1]])
```

Computing Principal Components

The A matrix

First, you should compute the `A` matrix, where A is $(X - \mu)'$. (Note the transpose)

Print this matrix below. It should have size 19×4.

In [52]:
```
# YOUR CODE GOES HERE
M = np.mean(X, axis=0)
A = (X - M).T
print("A = \n", A)
```

```
A =
 [[-2.75  0.25  0.25  2.25]
 [ 1.25  2.25 -2.75 -0.75]
 [ 2.   -4.    2.    0.  ]
 [-3.5   1.5   0.5   1.5 ]
 [ 3.5  -4.5  -0.5   1.5 ]
 [ 2.25  3.25 -1.75 -3.75]
 [ 1.    6.   -4.   -3.  ]
 [ 2.25  1.25 -1.75 -1.75]
 [-1.75  0.25  1.25  0.25]
 [ 0.75 -0.25  1.75 -2.25]
 [ 1.25 -2.75 -1.75  3.25]
 [ 0.5  -0.5  -3.5   3.5 ]
 [ 1.   -1.   -3.    3.  ]
 [ 1.75 -0.25 -2.25  0.75]
 [-1.5  -1.5   1.5   1.5 ]
 [-3.    1.    0.    2.  ]
 [ 0.5  -0.5   3.5  -3.5 ]
 [ 0.25 -3.75  3.25  0.25]
 [ 0.25 -1.75  2.25 -0.75]]
```

"Small" covariance matrix

By transposing $X - \mu$ to get A, now we can compute a smaller covariance matrix with $A'A$. Compute this matrix, `C`, below and print the result.

In [53]:
```
# YOUR CODE GOES HERE
C = A.T @ A
print("C = \n", C)
```

```
C =
 [[ 69.875 -18.875 -26.375 -24.625]
 [-18.875 121.375 -53.125 -49.375]
 [-26.375 -53.125  98.375 -18.875]
 [-24.625 -49.375 -18.875  92.875]]
```

Finding nonzero eigenvectors

Next, find the useful (nonzero) eigenvectors of `C`.

For validation purposes, there should be 3 useful eigenvectors, and the first one is `[-0.06628148 -0.79038331 0.47285044 0.38381435]`.

Keep these eigenvectors in a 4×3 array `e`.

In [54]:
```
# YOUR CODE GOES HERE
v,e = np.linalg.eig(C)
v,e = np.real(v), np.real(e)
id = np.argsort(-v)
v,e = v[idx],e[:, id]
nonzero_ind = np.where(np.abs(v) > 1e-10)[0]
v = v[nonzero_ind]
e = e[:,nonzero_ind]
print("The First useful eigenvectors is = ", e.T)
```

```
The First useful eigenvectors is =  [[-0.06628148 -0.79038331  0.47285044  0.38381435]
 [ 0.04124587 -0.06822502 -0.69123739  0.71821654]
 [-0.86249959  0.34733208  0.22046165  0.29470586]]
```

Calculating "eigenfaces"

Now, we have all we need to compute `U`, the matrix of eigenfaces.

$$\mathbf{U_i} = \mathbf{A e_i}$$

$$(19 \times 3) = (19 \times 4)(4 \times 3)$$

Compute and print U. Be sure to normalize your eigenvectors `e` before using the above equation.

In [55]:
```
# YOUR CODE GOES HERE
U = A @ e
U/= np.linalg.norm(U, axis=0)
e/= np.linalg.norm(e, axis=0)

print("Eigenfaces, U:\n",U)
```

```
Eigenfaces, U:
 [[ 0.07294372  0.12277459  0.33008441]
 [-0.26034151  0.11787331 -0.11677714]
 [ 0.29998485 -0.09606164 -0.27776956]
 [-0.01067529  0.04536213  0.42516696]
 [ 0.27653993  0.17530224 -0.44157072]
 [-0.37621372 -0.15082188 -0.23925816]
 [-0.59257956  0.02265222 -0.05657115]
 [-0.19897063 -0.0037123  -0.250194  ]
 [ 0.04569305 -0.07236581  0.20213547]
 [ 0.0084373  -0.25979087 -0.10504274]
 [ 0.18948616  0.35382298 -0.1518308 ]
 [ 0.00380575  0.46650428 -0.03585222]
 [ 0.03449119  0.40571147 -0.10256065]
 [-0.05241297  0.20419008 -0.19442141]
 [ 0.19396809  0.00756997  0.16057937]
 [ 0.01329023  0.11639359  0.36617258]
 [ 0.0508452  -0.45626561 -0.08985059]
 [ 0.3456779  -0.16842745 -0.07563409]
 [ 0.16171488 -0.18371276 -0.0569842 ]]
```

Projecting data into 3D

Now project your data into 3 dimensions with the formula:

$$\Omega = U^{\mathrm{T}} A$$

$$(3 \times 4) = (3 \times 19)(19 \times 4)$$

Call the projected data Ω "`W`". Print `W.T`

In [56]:
```
# YOUR CODE GOES HERE
W = U.T @ A
print('Projected data in 3 dimensions:\n',W.T)
```

```
Projected data in 3 dimensions:
 [[ -0.8782013    0.44099733  -8.3011616 ]
 [-10.47224127  -0.72945617   3.34291139]
 [  6.26506632  -7.39065157   2.12184196]
 [  5.08537624   7.67911041   2.83640825]]
```

Reconstructing data in 19-D

We can project the transformed data `W` back into the original 19-D space using:

$\Gamma_f = U\Omega + \Psi$
where:
$\Gamma_f =$ reconstructed data
$U =$ eigenfaces
$\Omega =$ Reduced data
$\Psi =$ Means

Do this, and compute the MSE between each reconstructed sample and corresponding original points. Report all 4 MSE values.

In [57]:
```
# YOUR CODE GOES HERE
Gammaf = U @ W + M.reshape(-1, 1)
org_data = X.T
MSE = np.mean((Gammaf - org_data) ** 2, axis = 0)

for i in range(4):
    print("MSE for sample %d: %e" %(i+1,MSE[i]))
```

```
MSE for sample 1: 6.705074e-31
MSE for sample 2: 2.002805e-30
MSE for sample 3: 3.258686e-30
MSE for sample 4: 1.308425e-30
```

2-D Reconstruction

What if we had only used the first 2 eigenvectors to compute the eigenfaces? Below, redo the earlier calculations, but use only two eigenfaces. Compute the 4 MSE values that you would get in this case.

(You should get an MSE of 3.626 for the first sample.)

In [59]:
```
# YOUR CODE GOES HERE
e = e[:,:2]
U = A @ e
```

```
U/= np.linalg.norm(U, axis=0)
W = U.T @ A
Gammaf = U @ W + M.reshape(-1,1)
MSE2 = np.mean((Gammaf - org_data) ** 2, axis = 0)

print("Using only 2 eigenvectors:")
for i in range(4):
    print("MSE for sample %d: %e" %(i+1,MSE2[i]))
```

```
Using only 2 eigenvectors:
MSE for sample 1: 3.626804e+00
MSE for sample 2: 5.881609e-01
MSE for sample 3: 2.369586e-01
MSE for sample 4: 4.234322e-01
```

Problem 1

Problem Description

In this problem you will use PCA and TSNE to apply dimensionality reduction to 64x64 images of signed distance fields (SDFs) on parts belonging to 8 different classes. Each class is topologicaly similar, with some variation in void size and shape. These signed distance fields are helpful in the prediction of internal stress fields in the parts. You will also apply KNN to predict the class of the part with the reduced space.

Fill out the notebook as instructed, making the requested plots and printing necessary values.

You are welcome to use any of the code provided in the lecture activities.

Summary of deliverables:

- 3x8 subplot visualization of the first 3 samples from each of the 8 classes
- Bar plot of the variance explained for the first 25 PCs and the number of PCs required to explain > 90% of the variance in the training data
- 4x8 subplot visualization of reconstructed samples using 3, 10, 50 and all PCs on the first sample from each of the 8 classes in the test set
- Test accuracy of KNN classifier trained on the 3D, 10D, and 50D PCA reduced feature spaces
- Plot of the 2D TSNE reduced feature space
- Test accuracy of the KNN classifier trained on the 2D TSNE reduced feature space
- Discussion questions 1 and 2

Imports and Utility Functions:

In [51]:
```
import numpy as np
import matplotlib.pyplot as plt
from scipy import io

from sklearn.decomposition import PCA
from sklearn.manifold import TSNE
```

```
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split

def dataLoader(filepath):
    # Load and flatten the SDF dataset
    mat = io.loadmat(filepath)
    data = []
    for i in range(800):
        sdf = mat["sdf"][i][0].T
        data.append(sdf.flatten())
    data = np.vstack(data)
    # Assign labels
    labels = np.repeat(np.arange(8), 100)
    return data, labels

def plot_sdf(data, ax = None, title = None):
    # If no axes, make them
    if ax is None:
        ax = plt.gca()
    # Reshape image data into square
    sdf = data.reshape(64,64)
    # Plot image, with bounds of the SDF values for the entire dataset
    ax.imshow(sdf, vmin=-0.31857, vmax=0.206349, cmap="jet")
    ax.axis('off')
    # If there is a title, add it
    if title:
        ax.set_title(title)
```

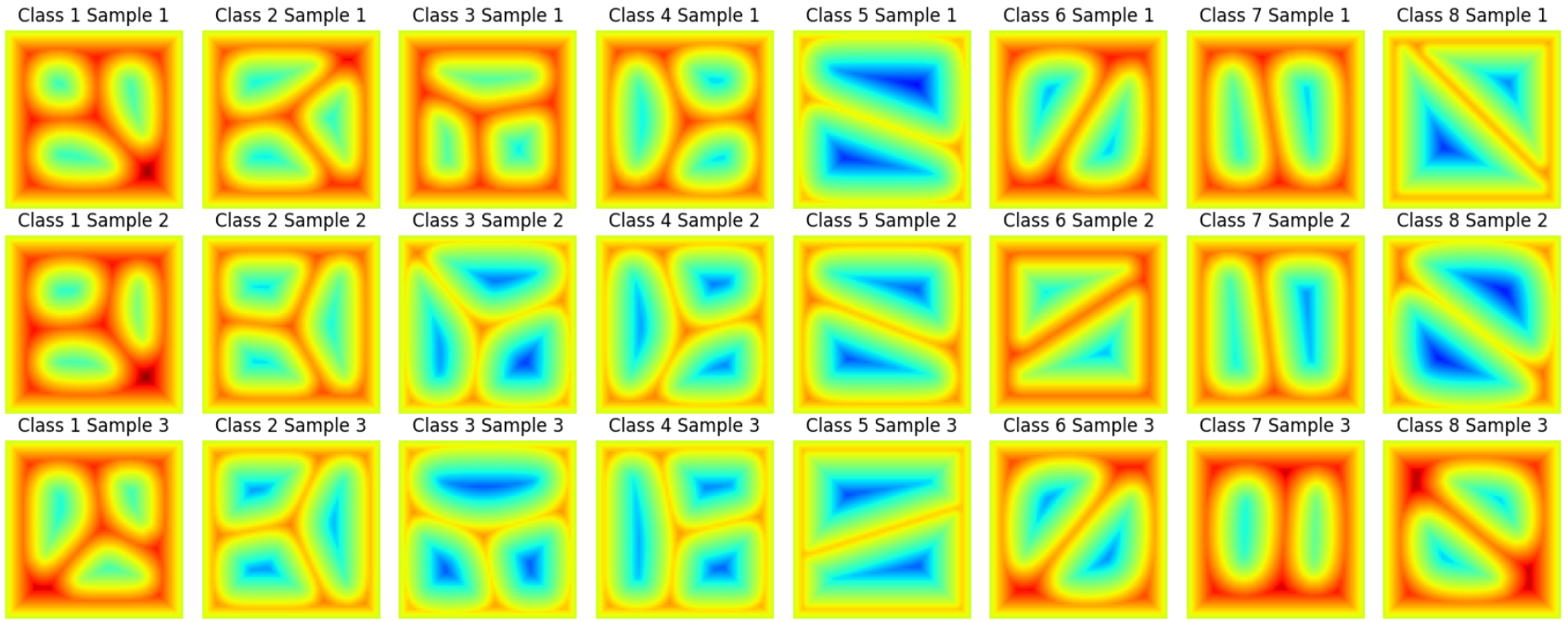
Visualization

Using the provided `dataLoader()` function, load the data and labels from `sdf_images.mat`. The returned data will contain 800 samples, with 4096 features. Then, using the provided `plot_sdf()` function, generate a 3x8 subplot figure containing visualizations of the first 3 SDFs in each class.

In [52]:
```
# YOUR CODE GOES HERE
file_path = r"C:\Users\srech\Downloads\data\sdf_images.mat"
data, labels = dataLoader(file_path)
fig, axes = plt.subplots(3, 8, figsize=(15, 6))
```

```
for i in range(8):
    class_indices = np.where(labels == i)[0][:3]
    for j, idx in enumerate(class_indices):
        plot_sdf(data[idx], ax=axes[j, i], title=f"Class {i+1} Sample {j+1}")

plt.tight_layout()
plt.show()
```



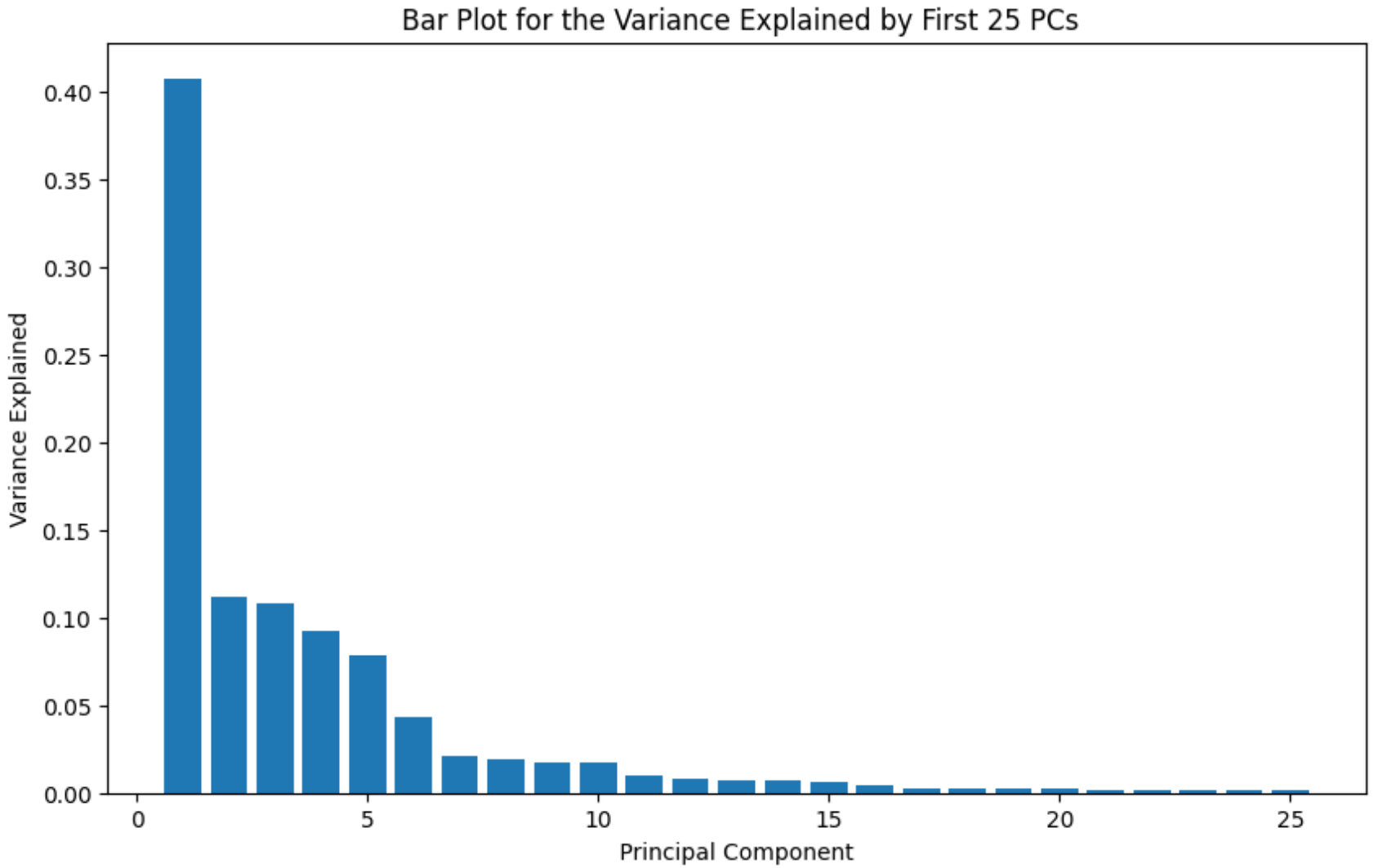
Explained Variance

Use `train_test_split()` to partition the data and labels into a training and test set with `test_size = 0.2` and `random_state = 0`. Then train a PCA model on the training data and generate a bar plot of the variance explained for the first 25 principal components. Determine the number of principal components required to explain > 90% of the variance in the training data.

In [54]:
```
# YOUR CODE GOES HERE
X_train, X_test, y_train, y_test = train_test_split(data, labels, test_size=0.2, random_state=0)
pca = PCA( n_components = 25)
```

```
pca.fit(X_train)
plt.figure(figsize = (10, 6))
plt.bar(range(1, 26), pca.explained_variance_ratio_)
plt.xlabel(' Principal Component ')
plt.ylabel(' Variance Explained ')
plt.title(' Bar Plot for the Variance Explained by First 25 PCs')
plt.show()

cum_var = np.cumsum(pca.explained_variance_ratio_)
n_components90 = np.where(cum_var > 0.9)[0][0] + 1
print(f"Number of principal components to explain > 90% variance: {n_components90}")
```




```
Number of principal components to explain > 90% variance: 9
```

PCA Reconstruction

Using the training data, generate 4 PCA models using 3, 10, 50, and all of the principal components. Use these models to transform the test data into the reduced space, and then reconstruct the data from the reduced space. Plot the reconstruction for each model,

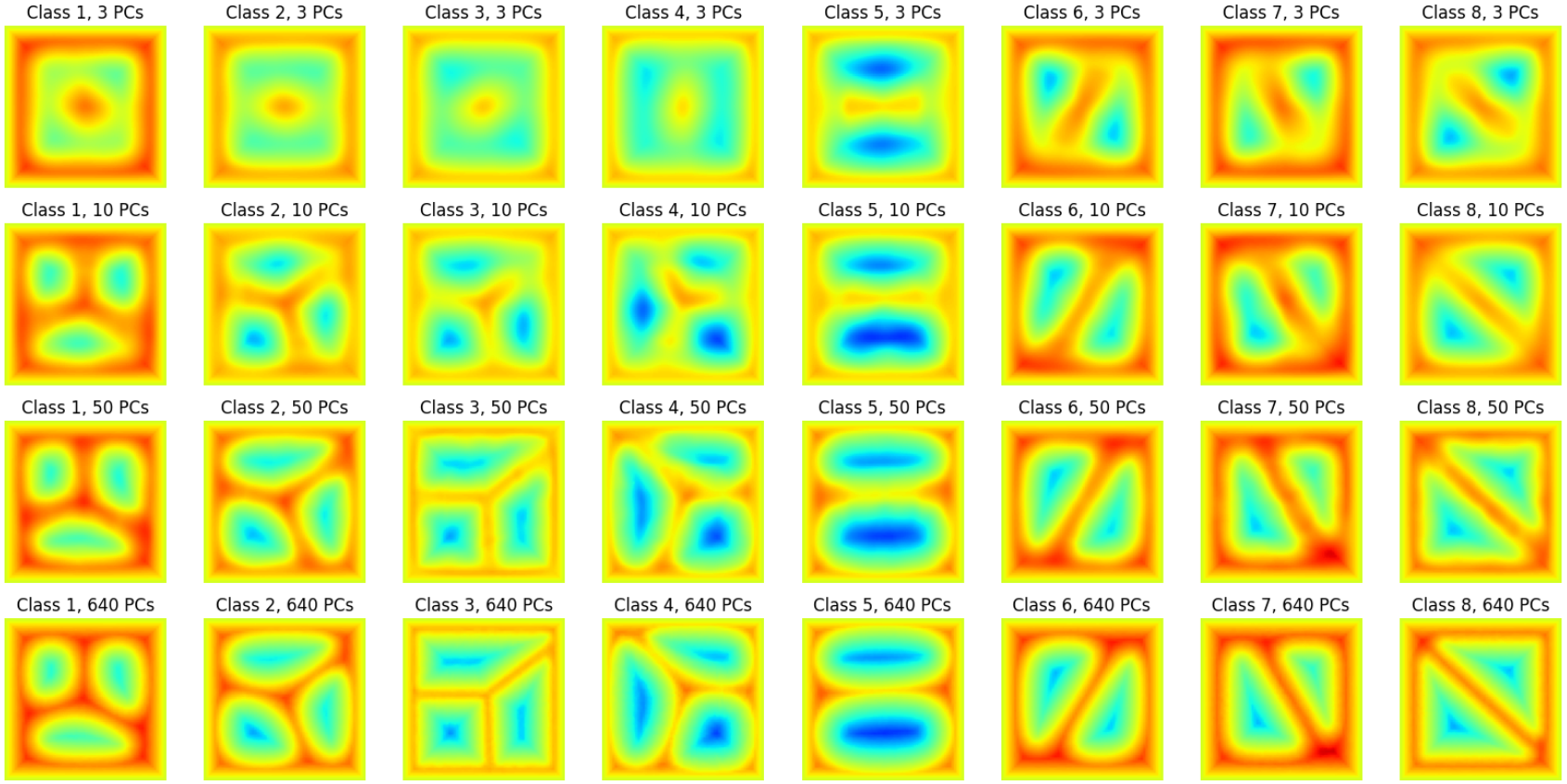
on the first occurence of each class in the test set. Your generated plot should be a 4x8 subplot figure, with each subplot title containing the class and the number of PCs used.

In [55]:
```
# YOUR CODE GOES HERE
n_components_all = min(X_train.shape)
n_components_list = [3, 10, 50, n_components_all]
fig, axes = plt.subplots(4, 8, figsize=(16, 8))

for i, n_components in enumerate(n_components_list):
    pca = PCA(n_components=n_components)
    pca.fit(X_train)
    X_test_pca = pca.transform(X_test)
    X_test_reconstructed = pca.inverse_transform(X_test_pca)

    for class_index in range(8):
        idx = next(j for j, label in enumerate(y_test) if label == class_index)
        ax = axes[i, class_index]
        plot_sdf(X_test_reconstructed[idx], ax, title=f"Class {class_index+1}, {n_components} PCs")

plt.tight_layout()
plt.show()
```



KNN on PCA Reduced Data

Now train a KNN classifier to predict the class of the 3D, 10D, and 50D PCA reduced data. You should train the KNN on the reduced training data, and report the prediction accuracy on the test set. You will also need to determine the `n_neighbors` parameter for your KNN classifier that gives good results.

In [56]:
```
# YOUR CODE GOES HERE
from sklearn.model_selection import cross_val_score
num_components_list = [3, 10, 50]

for num_components in num_components_list:
    pca = PCA(n_components=num_components)
    X_train_pca = pca.fit_transform(X_train)
```

```
    knn = KNeighborsClassifier()
    k_values = list(range(1, 21))
    accuracies = []
    for k in k_values:
        knn.n_neighbors = k
        scores = cross_val_score(knn, X_train_pca, y_train, cv=5)  # Corrected variable names
        accuracies.append(np.mean(scores))

    optimal_k = k_values[np.argmax(accuracies)]
    knn.n_neighbors = optimal_k
    knn.fit(X_train_pca, y_train)
    X_test_pca = pca.transform(X_test)
    predictions = knn.predict(X_test_pca)
    accuracy = np.mean(predictions == y_test)

    print(f"The Accuracy for {num_components}-D PCA reduced data with {optimal_k} neighbors is = {accuracy*100:.2f}%"
```

```
The Accuracy for 3-D PCA reduced data with 15 neighbors is = 70.62%
The Accuracy for 10-D PCA reduced data with 14 neighbors is = 91.25%
The Accuracy for 50-D PCA reduced data with 14 neighbors is = 92.50%
```

TSNE Visualization

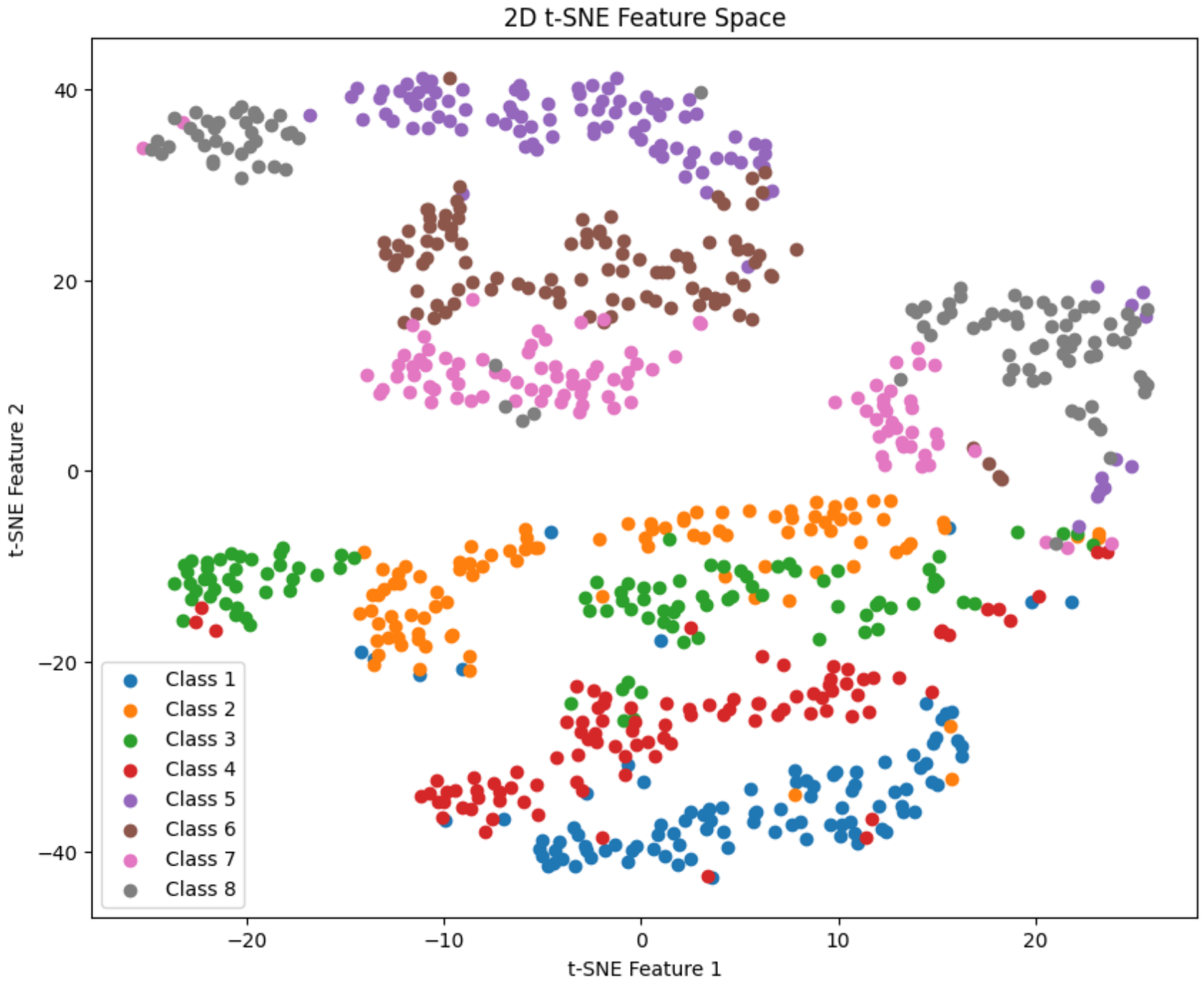
First reduced the full dataset to 50D using PCA, and then further reduced the data to 2D using TSNE. Plot the 2D reduced feature space with a scatter plot, coloring each point according to its class.

In [59]:
```
# YOUR CODE GOES HERE
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.manifold import TSNE

scaler = StandardScaler()
data_scaled = scaler.fit_transform(data)
pca_50 = PCA(n_components=50)
data_pca_50 = pca_50.fit_transform(data_scaled)
tsne = TSNE(n_components=2, random_state=0)
data_tsne_2d = tsne.fit_transform(data_pca_50)
plt.figure(figsize=(10, 8))
for class_index in range(8):
    indices = labels == class_index
```

```
    plt.scatter(data_tsne_2d[indices, 0], data_tsne_2d[indices, 1], label=f"Class {class_index+1}")

plt.xlabel('t-SNE Feature 1')
plt.ylabel('t-SNE Feature 2')
plt.title('2D t-SNE Feature Space')
plt.legend()
plt.show()
```


2D t-SNE Feature Space
t-SNE Feature 2
t-SNE Feature 1
40
20
0
−20
−40
−20
−10
0
10
20
Class 1
Class 2
Class 3
Class 4
Class 5
Class 6
Class 7
Class 8

KNN on PCA/TSNE Reduced Data

Using the same 2D PCA/TSNE data, split the data into train and test data and labels using `train_test_split` with a `random_state = 0` parameter so you have the same train/test partition as before. Then, train a KNN on this 2D feature space with the training set, and report the KNN classifier accuracy on the test set. Again, you will need to determine the `n_neighbors` parameter in the KNN classifier that gives good results.

In [60]:
```
# YOUR CODE GOES HERE
from sklearn.decomposition import PCA
from sklearn.manifold import TSNE
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score

pca_50 = PCA(n_components=50)
data_50d = pca_50.fit_transform(data)
tsne = TSNE(n_components=2, random_state=0)
data_2d = tsne.fit_transform(data_50d)
X_train_tsne, X_test_tsne, y_train_tsne, y_test_tsne = train_test_split(data_2d, labels, test_size=0.2, random_state=

def train_and_evaluate_knn_2d(X_train, y_train, X_test, y_test, n_neighbors):
    knn = KNeighborsClassifier(n_neighbors=n_neighbors)
    knn.fit(X_train, y_train)
    y_pred = knn.predict(X_test)
    accuracy = accuracy_score(y_test, y_pred)
    return accuracy
# Change and play around
n_neighbors = 6
accuracy_tsne = train_and_evaluate_knn_2d(X_train_tsne, y_train_tsne, X_test_tsne, y_test_tsne, n_neighbors)
print(f"Accuracy of KNN classifier on 2D TSNE reduced data: {accuracy_tsne:.2f}")
```

```
Accuracy of KNN classifier on 2D TSNE reduced data: 0.91
```

Discussion

1. Discuss how the number of principal components relates to the quality of reconstruction of the data. Using all of the principal components, should there be any error in the reconstruction of a sample from the training data? What about in the

reconstruction of an unseen sample from the testing data?
2. Discuss how you determined `k`, the number of neighbors in your KNN models. Why do we perform dimensionality reduction to our data before feeding it to our KNN classifier?

Your response goes here

1. In Principal Component Analysis (PCA), the selected number of principal components significantly impacts how accurately the original data can be reconstructed. Fewer components lead to greater reconstruction errors, as they capture only a portion of the data's variance. This results in a compressed representation that may miss finer details. However, increasing the number of components enhances the accuracy of reconstruction, with all components offering a complete and error-free reconstruction of the training data. This perfect reconstruction, though, may not apply to unseen test data, as PCA is optimized based on the training dataset, and may not accurately reflect the variance in new data. This discrepancy highlights the importance of a model's generalizability. In practical applications, a balanced number of components, like 20, can often provide a sufficiently accurate reconstruction while maintaining computational efficiency. For instance, using 10 PCs yields a 91.25% accuracy, which only marginally improves to 92.50% with 50 PCs, suggesting that beyond a certain point, adding more components offers diminishing returns in accuracy.

2. To determine the optimal 'k' for KNN models, we used cross-validation, training the classifier with various 'k' values and selecting the one that maximized accuracy. Dimensionality reduction, like PCA, is essential before applying KNN to reduce computational complexity and avoid the "curse of dimensionality," which affects distance-based methods in high-dimensional spaces. This reduction makes distances more meaningful, improves classifier performance, filters out noise, and enhances data interpretability.