# Deep Learning (IST, 2021-22)

# Practical 7: PCA and Autoencoders

Gonçalo Faria, André Martins, Andreas Wichert, Luis Sá-Couto

## Question 1

Consider the following training data:

$$\boldsymbol{x}^{(1)} = \begin{bmatrix} 2.5 \\ 2.4 \end{bmatrix}, \quad \boldsymbol{x}^{(2)} = \begin{bmatrix} 0.5 \\ 0.7 \end{bmatrix}, \quad \boldsymbol{x}^{(3)} = \begin{bmatrix} 2.2 \\ 2.9 \end{bmatrix}, \quad \boldsymbol{x}^{(4)} = \begin{bmatrix} 1.9 \\ 2.2 \end{bmatrix}$$

1. Find the set of orthogonal directions that maximize the variance of the training data. (Hint: perform principal component analysis.)

**Solution:**

The set of orthogonal directions that maximize the variance of the training data is the covariance matrix's eigenvectors. We first compute the training data's covariance matrix $\boldsymbol{C}$ as follows:

$$\mu = \frac{1}{4}\left(\begin{bmatrix} 2.5 \\ 2.4 \end{bmatrix} + \begin{bmatrix} 0.5 \\ 0.7 \end{bmatrix} + \begin{bmatrix} 2.2 \\ 2.9 \end{bmatrix} + \begin{bmatrix} 1.9 \\ 2.2 \end{bmatrix}\right) = \begin{bmatrix} 1.775 \\ 2.05 \end{bmatrix}$$

$$\boldsymbol{z}^{(1)} = \begin{bmatrix} 2.5 \\ 2.4 \end{bmatrix} - \begin{bmatrix} 1.775 \\ 2.05 \end{bmatrix} = \begin{bmatrix} 0.725 \\ 0.35 \end{bmatrix}$$

$$\boldsymbol{z}^{(2)} = \begin{bmatrix} 0.5 \\ 0.7 \end{bmatrix} - \begin{bmatrix} 1.775 \\ 2.05 \end{bmatrix} = \begin{bmatrix} -1.275 \\ -1.35 \end{bmatrix}$$

$$\boldsymbol{z}^{(3)} = \begin{bmatrix} 2.2 \\ 2.9 \end{bmatrix} - \begin{bmatrix} 1.775 \\ 2.05 \end{bmatrix} = \begin{bmatrix} 0.425 \\ 0.85 \end{bmatrix}$$

$$\boldsymbol{z}^{(4)} = \begin{bmatrix} 1.9 \\ 2.2 \end{bmatrix} - \begin{bmatrix} 1.775 \\ 2.05 \end{bmatrix} = \begin{bmatrix} 0.125 \\ 0.15 \end{bmatrix}$$

$$\boldsymbol{Z} = \begin{bmatrix} \text{——}\boldsymbol{z}^{(1)}\text{——} \\ \text{——}\boldsymbol{z}^{(2)}\text{——} \\ \text{——}\boldsymbol{z}^{(3)}\text{——} \\ \text{——}\boldsymbol{z}^{(4)}\text{——} \end{bmatrix} = \begin{bmatrix} 0.725 & 0.35 \\ -1.275 & -1.35 \\ 0.425 & 0.85 \\ 0.125 & 0.15 \end{bmatrix}$$

$$\boldsymbol{C} = \boldsymbol{Z}^\top \boldsymbol{Z} = \begin{bmatrix} 2.3475 & 2.355 \\ 2.355 & 2.69 \end{bmatrix}$$

To obtain the eigenvectors, we compute $\boldsymbol{C}$'s eigenvalues as follows:

$$\det\Big(\boldsymbol{C} - \lambda \boldsymbol{I}\Big) = 0 \Leftrightarrow$$

$$\det\Big(\begin{bmatrix} 2.3475 & 2.355 \\ 2.355 & 2.69 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\Big) = 0 \Leftrightarrow$$

$$(2.3475 - \lambda)(2.69 - \lambda) - 2.355^2 = 0$$

Solving this equation gives :

$$\lambda = 0.15753176 \quad \vee \quad \lambda = 4.87996824$$

If we now give special symbols for each of these solutions, particularly :

$$\lambda_1 = 0.15753176, \quad \lambda_2 = 4.87996824$$

We can compute the eigenvectors associated with these eigenvalues by solving the following two systems of linear equations:

$$\begin{bmatrix} 2.3475 - \lambda_1 & 2.355 \\ 2.355 & 2.69 - \lambda_1 \end{bmatrix} u_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \text{and} \quad \begin{bmatrix} 2.3475 - \lambda_2 & 2.355 \\ 2.355 & 2.69 - \lambda_2 \end{bmatrix} u_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Note that each of these equations have infinitely many solutions (e.g., if $\tilde{u}_1$ is a particular solution of the first equation, then any scaled version $\lambda\tilde{u}_1$ with $\lambda \in \mathbb{R}$ is also a solution). Therefore, we cannot use simple Gauss elimination, but we can assume without loss of generality that $\tilde{u}_1$ and $\tilde{u}_2$ are of the form $(\alpha, 1)$ and solve the equation for $\alpha$. This gives respectively:

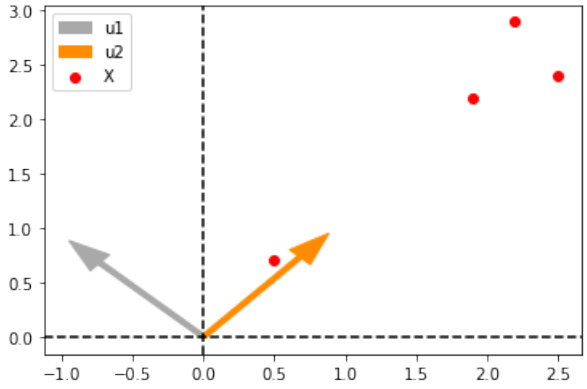
$$\tilde{u}_1 = \begin{bmatrix} -1.07535806 \\ 1 \end{bmatrix}, \quad \tilde{u}_2 = \begin{bmatrix} 0.929922817 \\ 1 \end{bmatrix}$$

Since $\lambda\tilde{u}_1$ and $\lambda\tilde{u}_2$ are also solutions for any $\lambda \in \mathbb{R}$, it is common to normalize these vectors, i.e., to return $\tilde{u}_1/\|\tilde{u}_1\|$ and $\tilde{u}_1/\|\tilde{u}_1\|$, which gives :

$$u_1 = \begin{bmatrix} -0.73229984 \\ 0.68098233 \end{bmatrix}, \quad u_2 = \begin{bmatrix} 0.68098233 \\ 0.73229984 \end{bmatrix}$$

2. Draw a Cartesian plane containing the training data in the original coordinates and the vectors of principal components.

**Solution:**




3. What is the first principal component?

   **Solution:** The first principle component is the eigenvector associated with the largest eigenvalue. In this case, it is the eigenvector $u_2$.

4. Reduce the dimensionality of the training data by mapping the points onto the principal component

   **Solution:**

$$\boldsymbol{Y} = \boldsymbol{X}u_2^\top = \begin{bmatrix} \text{——}\boldsymbol{x}^{(1)}\text{——} \\ \text{——}\boldsymbol{x}^{(2)}\text{——} \\ \text{——}\boldsymbol{x}^{(3)}\text{——} \\ \text{——}\boldsymbol{x}^{(4)}\text{——} \end{bmatrix} \begin{bmatrix} 0.68098233 \\ 0.73229984 \end{bmatrix} = \begin{bmatrix} 3.45997546 \\ 0.85310106 \\ 3.62183068 \\ 2.90492609 \end{bmatrix}$$

5. Find the orthogonal projection of the data onto the first principal component's subspace in the original coordinates.

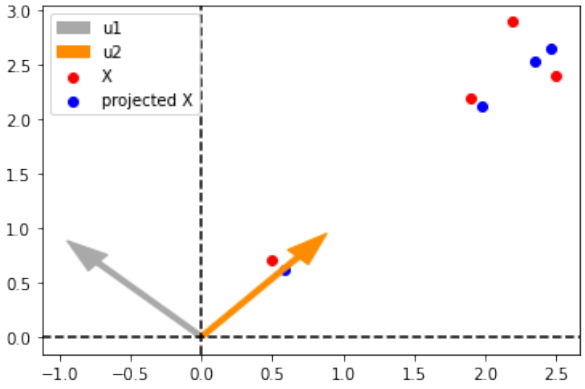
   **Solution:**

$$P = u_2 u_2^\top = \begin{bmatrix} 0.68098233 \\ 0.73229984 \end{bmatrix} \begin{bmatrix} 0.680982330.73229984 \end{bmatrix} = \begin{bmatrix} 0.46373694 & 0.49868326 \\ 0.49868326 & 0.53626306 \end{bmatrix}$$

6. Draw the projected training data on the previous Cartesian plane.

   **Solution:**

$$\hat{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{P} = \begin{bmatrix} \text{——}\boldsymbol{x}^{(1)}\text{——} \\ \text{——}\boldsymbol{x}^{(2)}\text{——} \\ \text{——}\boldsymbol{x}^{(3)}\text{——} \\ \text{——}\boldsymbol{x}^{(4)}\text{——} \end{bmatrix} \begin{bmatrix} 0.46373694 & 0.49868326 \\ 0.49868326 & 0.53626306 \end{bmatrix} = \begin{bmatrix} 2.35618216 & 2.53373949 \\ 0.58094675 & 0.62472577 \\ 2.46640271 & 2.65226604 \\ 1.97820335 & 2.12727692 \end{bmatrix}$$




7. What is the mean squared error of the projected training data?

**Solution:**

$$\text{MSE}(\boldsymbol{X}, \hat{\boldsymbol{X}}) = \frac{1}{4}\sum_{i=1}^{4} \|\boldsymbol{x}^{(i)} - \hat{\boldsymbol{x}}^{(i)}\|^2 = 0.04863383382534045$$

## Question 2

Now it's time to run PCA on real data.

1. Load the UCI handwritten digits dataset using `scikit-learn`:

```
from sklearn.preprocessing import MinMaxScaler
from sklearn.model_selection import train_test_split

from sklearn.datasets import load_digits
data = load_digits()

X, y = data.data, data.target
# normalize images.
X = MinMaxScaler().fit_transform(X)

noise = rng.normal(scale=0.25, size=X_train.shape)
X_train_noisy = X_train + noise
```

This is a dataset containing 1797 8x8 input images of digits, each corresponding to one out of 10 output classes. You can print the dataset description and visualize some input examples with:

```
def plot_digits(X, title):
    """Small helper function to plot 100 digits."""
    fig, axs = plt.subplots(nrows=10, ncols=10, figsize=(8, 8))
    for img, ax in zip(X, axs.ravel()):
        ax.imshow(img.reshape((8, 8)), cmap="Greys")
        ax.axis("off")
    fig.suptitle(title, fontsize=24)
```

Randomly split this data into training (80%) and test (20%) partitions. This can be done with:

```
X_train, X_test, y_train, y_test = train_test_split(
    X, y, stratify=y, random_state=0, train_size=1697, test_size=100
)
```

2. Apply independent Gaussian noise to each of the pixels $\epsilon \sim \mathcal{N}(0, 0.25)$ of every image. Plot the resulting images using `matplotlib`.

```
train_noise = 0.25 * np.random.randn(*X_train.shape)
```

3. Run your implementation of PCA on this dataset. Consider different numbers of principal components, and calculate the corresponding orthogonal projection of the images onto the first principal component's subspace in the original coordinate space. Can you recover the uncorrupted digits?

## Question 3

Now it's time to train an autoencoder, as depicted in Figure 1, on real data. We will use the same data as the previous exercise.

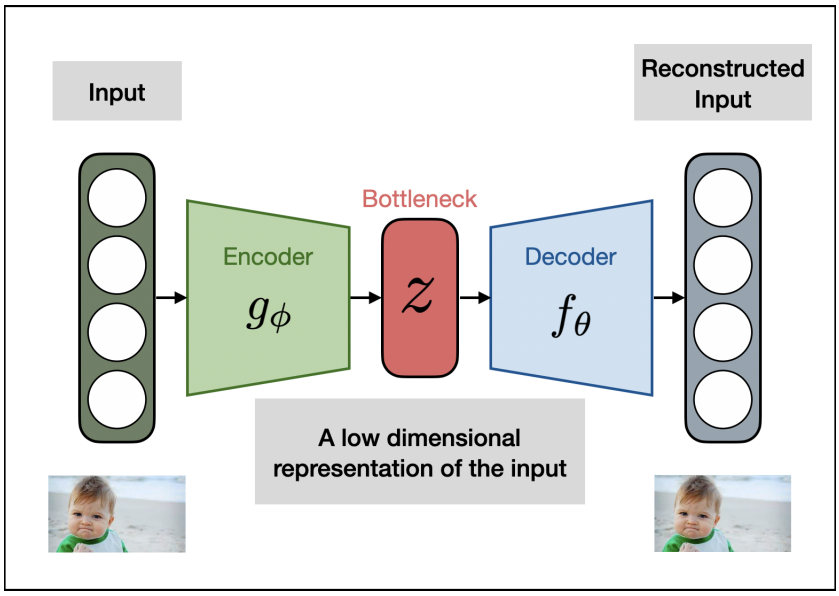



Figura 1: Diagram of an autoencoder.

1. Run your implementation of autoencoders with Linear activations on this dataset. Consider different optimizers and hyper-parameters (particularly bottleneck size).

   Can you see similarities with PCA?

2. Run your implementation of autoencoders with non-linear activations on this dataset. Consider different optimizers and hyper-parameters (particularly bottleneck size). Plot the reconstructed digits on the noisy test set.

3. Rerun your implementation of autoencoders with non-linear activations on this dataset with a bottleneck size of 3. Plot for each test set image its 3-dimensional representation using your model, and color each point according to the image's class.