Feature Extraction

Ruwen Qin
Stony Brook University

September 4, 2025

Data used in this module:

- California house prices dataset from sklearn.datasets
- Optical recognition of handwritten digits dataset from sklearn.datasets

Python notebook used in this chapter

- Feature_Extraction.ipynb

1 Background

Considering a system characterized by N attributes. We collect a sample data of the attributes, which contains M observations. Denote the sample data as $\mathbf{X} \in \mathbb{R}^{M \times N}$. If N is large, we hope to characterize the system with fewer but informative features. In some cases N is not large, but some attributes may be correlated with each other. We hope to find features that independently characterize the system from unique perspectives.

In this module, we will study principal component analysis and then briefly introduce a few more ML methods for feature extraction and dimension reduction.

2 Principal Component Analysis (PCA)

Define $\mathcal{P}_d$ as the set of N-dimensional rank-d orthogonal projection matrix[1]. We can project the zero-centered sample data $\widetilde{\mathbf{X}}(= \mathbf{X} - \overline{\mathbf{X}})$ onto the d-dimensional linear sub-space using a projection matrix $\mathbf{P} \in \mathcal{P}_d$. We hope to minimize the reconstruction error; that is, the sum of L_2 distances between the original data $\widetilde{\mathbf{X}}$ and the projected data $\widetilde{\mathbf{X}}\mathbf{P}$:

$$\min_{\mathbf{P} \in \mathcal{P}_d} \|\widetilde{\mathbf{X}}\mathbf{P} - \widetilde{\mathbf{X}}\|_F^2, \tag{1}$$

where $\|\cdot\|_F$ is the Frobenius norm for matrices.

The *Principal Component Analysis* (PCA) [1] shows that the solution to the minimization problem in (1), $\mathbf{P}^*$, contains d columns that are top d eigenvectors of $\mathbf{X}$'s covariance matrix Σ:

$$\Sigma = \frac{(\mathbf{X} - \overline{\mathbf{X}})^{\mathrm{T}}(\mathbf{X} - \overline{\mathbf{X}})}{N-1} = \frac{\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{X}}}{N-1} \tag{2}$$

with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_d$[2].

2.1 Principal Components

Principal components (PCs) are normalized linear combinations of the feature vectors:

$$\mathbf{C}_j = \widetilde{\mathbf{X}}\mathbf{p}_j \tag{3}$$

The variance of the jth PC is

$$\mathrm{Var}[\mathbf{C}_j] = \mathbf{p}_j^T \Sigma \mathbf{p}_j, \tag{4}$$

which is the eigenvalue λ_j corresponding to the eigenvector $\mathbf{p}_j$. Therefore,

$$\frac{\lambda_j}{\sum_{i=1}^{d} \lambda_i} \tag{5}$$

is the proportion of variance explained by the jth PC and

$$\frac{\sum_{i=1}^{j} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \tag{6}$$

is the proportion of variance explained by the first j PCs.

[1] an $N \times N$ square matrix that can be written as UU^{T} where $U \in R^{N \times d}$ has d orthogonal columns.

[2] In linear algebra, we also know that the eigenvectors of $\mathbf{X}$'s co-variance matrix are the right singular vectors of $\widetilde{\mathbf{X}}$ and the eignvalues of $\mathbf{X}$'s co-variance matrix are squares of $\widetilde{\mathbf{X}}$'s positive eigenvalues divided by (N-1)

2.2 Derivation of PCA Algorithm

To minimize the reconstruction loss

$$\begin{aligned}\|\widetilde{\mathbf{X}}\mathbf{P}-\widetilde{\mathbf{X}}\|_F^2 &= \mathrm{Tr}((\widetilde{\mathbf{X}}\mathbf{P}-\widetilde{\mathbf{X}})(\widetilde{\mathbf{X}}\mathbf{P}-\widetilde{\mathbf{X}})^{\mathrm{T}}) \\ &= \mathrm{Tr}(\widetilde{\mathbf{X}}\mathbf{P}\mathbf{P}^{\mathrm{T}}\widetilde{\mathbf{X}}^{\mathrm{T}}-2\widetilde{\mathbf{X}}\mathbf{P}\widetilde{\mathbf{X}}^{\mathrm{T}}+\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^{\mathrm{T}}) \\ &= -\mathrm{Tr}(\widetilde{\mathbf{X}}\mathbf{P}\widetilde{\mathbf{X}}^{\mathrm{T}})+\mathrm{Tr}(\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^{\mathrm{T}})\end{aligned} \tag{7}$$

is to maximize $\mathrm{Tr}(\widetilde{\mathbf{X}}\mathbf{P}\widetilde{\mathbf{X}}^{\mathrm{T}})$[3].

Because $\mathbf{P}$ is an N-dimensional rank-d orthogonal projection matrix, $\mathbf{P}=\mathbf{U}\mathbf{U}^{\mathrm{T}}$, where $\mathbf{U}\in\mathbb{R}^{N\times d}$ containing orthogonal columns. Therefore,

$$\mathrm{Tr}(\widetilde{\mathbf{X}}\mathbf{P}\widetilde{\mathbf{X}}^{\mathrm{T}})=\mathrm{Tr}(\widetilde{\mathbf{X}}\mathbf{U}\mathbf{U}^{\mathrm{T}}\widetilde{\mathbf{X}}^{\mathrm{T}})=\mathrm{Tr}(\mathbf{U}^{\mathrm{T}}\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{X}}\mathbf{U})^{4}=\sum_{j=1}^{d}\mathbf{u}_j^{\mathrm{T}}\widetilde{\mathbf{X}}^{\mathrm{T}}\widetilde{\mathbf{X}}\mathbf{u}_j, \tag{8}$$

where $\mathbf{u}_j$ is the jth column of $\mathbf{U}$.

Therefore, we solve the following optimization problem:

$$\begin{aligned}&\max_{\mathbf{u}_1,\ldots,\mathbf{u}_d}\sum_{j=1}^{d}\mathbf{u}_j^{\mathrm{T}}\Sigma\mathbf{u}_j \\ &\text{subject to:}\quad \mathbf{u}_j^{\mathrm{T}}\mathbf{u}_j=1,\quad \text{for } j=1,\ldots,d,\end{aligned} \tag{9}$$

which is equivalent to

$$\max_{\mathbf{u}_1,\ldots,\mathbf{u}_d,\lambda_1,\ldots,\lambda_d}L=\sum_{j=1}^{d}\mathbf{u}_j^{\mathrm{T}}\Sigma\mathbf{u}_j-\sum_{j=1}^{d}\lambda_j(\mathbf{u}_j^{\mathrm{T}}\mathbf{u}_j-1). \tag{10}$$

This quadratic optimization problem is solved by

$$\begin{aligned}\frac{\partial L}{\partial \mathbf{d}_j}&=\Sigma\mathbf{u}_j-\lambda_j\mathbf{u}_j=0\quad \text{for } j=1,\ldots,d. \\ \frac{\partial L}{\partial \lambda_j}&=1-\mathbf{u}_j^{\mathrm{T}}\mathbf{u}_j=0\quad \text{for } j=1,\ldots,d.\end{aligned} \tag{11}$$

That is, λ_j's are eigenvalues of Σ and $\mathbf{u}_j$'s are the corresponding eigenvectors.

Apparently, the objective function is maximized by selecting the d largest eigenvalues and retaining the corresponding eigenvectors.

3 An Example with Two Attributes

3.1 The Data

Let's consider a system with only two attributes (i.e., N=2). We have a sample data of size 1,000 (i.e. M=1,000), $\mathbf{X}\in\mathbb{R}^{1000\times 2}$.

The sample means of the two features are:

$$\overline{\mathbf{X}}=[\overline{x}_1,\overline{x}_2]=[-0.0366,-0.0047]. \tag{12}$$

The sample covariance matrix is

$$\mathbf{\Sigma}=\frac{(\mathbf{X}-\overline{\mathbf{X}})^{\mathrm{T}}(\mathbf{X}-\overline{\mathbf{X}})}{N-1}=\begin{bmatrix}0.7922 & 0.4049\\ 0.4049 & 0.3207\end{bmatrix}. \tag{13}$$

[3]Since $\mathbf{P}$ is an orthogonal projection matrix, $\mathbf{P}=\mathbf{P}^{\mathrm{T}}$ and $\mathbf{P}^2=\mathbf{P}$. $\widetilde{\mathbf{X}}\widetilde{\mathbf{X}}^{\mathrm{T}}$ is a constant.
[4]Tr(AB)=Tr(BA) if matrix multiplication AB and BA both exist.

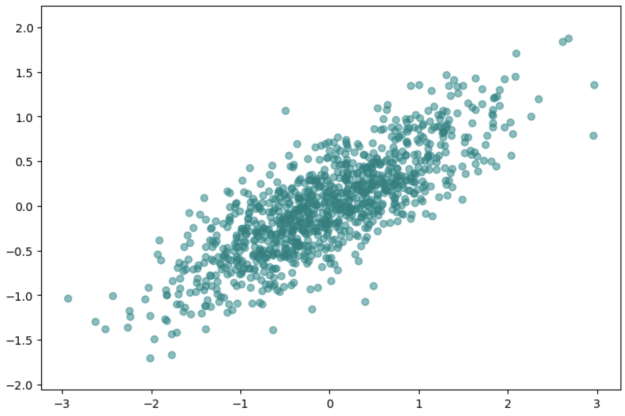


Figure 1: A sample of two-feature data in size 1000

The correlation coefficient is

$$r = \frac{0.4049}{\sqrt{0.7922}\sqrt{0.3207}} = 0.803, \tag{14}$$

which means the two features are strongly correlated with each other.

3.2 Eigen Decomposition of Covariance Matrix

We perform the eigen decomposition for $\mathbf{X}$'s covariance matrix. The eigenvectors are

$$\mathbf{V} = \begin{bmatrix} 0.8670 & -0.4984 \\ 0.4984 & 0.8670 \end{bmatrix} \approx \begin{bmatrix} \cos(30) & -\sin(30) \\ \sin(30) & \cos(30) \end{bmatrix}. \tag{15}$$

The corresponding eigenvalues in a diagonal matrix are

$$\mathbf{\Lambda} = \begin{bmatrix} 1.0249 & 0 \\ 0 & 0.0879 \end{bmatrix}. \tag{16}$$

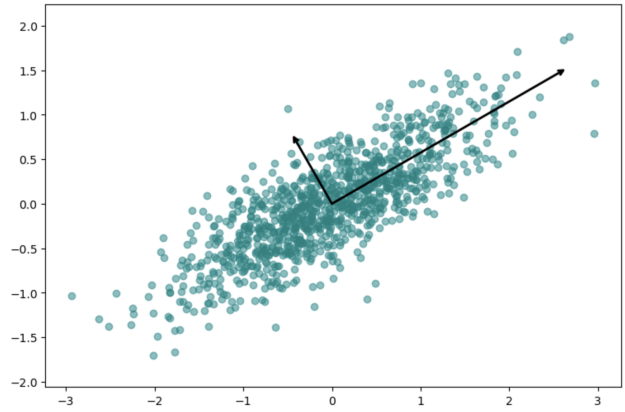


Figure 2: A sample of two-feature data in size 1000

The eigenvectors define two perpendicular directions. As Figure 2 shows, the first direction,

$$\mathbf{v}_1 = \begin{bmatrix} 0.8670 \\ 0.4984 \end{bmatrix} \tag{17}$$

contains the most variation of the data. The corresponding eigenvalue $\lambda_1 = 1.0249$ measures the amount of variance in this direction.

The second direction is defined by the second eigenvector,

$$\mathbf{v}_2 = \begin{bmatrix} -0.4984 \\ 0.8670 \end{bmatrix} \tag{18}$$

The variance in this direction is measured by the eigenvalue $\lambda_2 = 0.0809$.

3.3 PCA with Two Principal Components

Let's perform PCA on the zero-centered data $\widetilde{\mathbf{X}}$. The two projection vectors, which are the eigenvectors of the covariance matrix of data $\mathbf{X}$, defines the projection matrix:

$$\mathbf{P} = [\mathbf{p}_1 \quad \mathbf{p}_2] = \begin{bmatrix} 0.8670 & -0.4984 \\ 0.4984 & 0.8670 \end{bmatrix}. \tag{19}$$

They respectively explain certain amount of variances, measured by their eigenvalues:

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} = \begin{bmatrix} 1.0249 & 0 \\ 0 & 0.0879 \end{bmatrix}. \tag{20}$$

You must have found that

- the principal components are normalized linear combinations of $\mathbf{X}$'s features using the eigenvectors of $\mathbf{X}$'s sample covariance matrix
- the variances that principal components explain are the eigenvalues of $\mathbf{X}$'s sample covariance matrix.

The variance explained by the first principal component is:

$$\frac{1.0249}{1.0249 + 0.0879} = 92.1\%. \tag{21}$$

Therefore, using the first principal component as a feature seems also to be sufficient.

4 An Example of Feature Extraction using PCA

4.1 California Housing Dataset

Let's practice the feature extraction using PCA. We use the California Housing dataset provided by 'sklearn.dataset'.

```
.. _california_housing_dataset:

California Housing dataset
--------------------------

**Data Set Characteristics:**

    :Number of Instances: 20640

    :Number of Attributes: 8 numeric, predictive attributes and the target

    :Attribute Information:
        - MedInc        median income in block group
        - HouseAge      median house age in block group
        - AveRooms      average number of rooms per household
        - AveBedrms     average number of bedrooms per household
        - Population    block group population
        - AveOccup      average number of household members
        - Latitude      block group latitude
        - Longitude     block group longitude

    :Missing Attribute Values: None

This dataset was obtained from the StatLib repository.
```

```
https://www.dcc.fc.up.pt/~ltorgo/Regression/cal_housing.html

The target variable is the median house value for California districts,  expressed in
    hundreds of thousands of dollars ($100,000).

This dataset was derived from the 1990 U.S. census, using one row per census
block group.  A block group is the smallest geographical unit for which the U.S.  Census
    Bureau publishes sample data (a block group typically has a population of 600 to 3,000
    people).

A household is a group of people residing within a home. Since the average number of rooms
    and bedrooms in this dataset are provided per household, these columns may take
    surprisingly large values for block groups with few households and many empty houses,
    such as vacation resorts.

It can be downloaded/loaded using the
:func:`sklearn.datasets.fetch_california_housing` function.

.. topic:: References

    - Pace, R. Kelley and Ronald Barry, Sparse Spatial Autoregressions,
      Statistics and Probability Letters, 33 (1997) 291-297
```

4.2 Data Normalization

Denote the data as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$. When we check the first few observations of the dataset, we found that those attributes have very different scales.

```
   MedInc  HouseAge  AveRooms  AveBedrms  Population  AveOccup  Latitude  Longitude
0  8.3252      41.0  6.984127   1.023810       322.0  2.555556     37.88    -122.23
1  8.3014      21.0  6.238137   0.971880      2401.0  2.109842     37.86    -122.22
2  7.2574      52.0  8.288136   1.073446       496.0  2.802260     37.85    -122.24
3  5.6431      52.0  5.817352   1.073059       558.0  2.547945     37.85    -122.25
4  3.8462      52.0  6.281853   1.081081       565.0  2.181467     37.85    -122.25
```

To eliminate the impact of varied scales of features, we normalize the data through centering (subtracting the sample mean) and scaling (dividing the sample standard deviation):

$$\widetilde{\mathbf{x}_j} = \frac{\mathbf{x}_j - \overline{x}_j}{s_j} \tag{22}$$

for $j = 1, \ldots, N$ before performing PCA. Here,

$$\overline{x}_j = \frac{\sum_{i=1}^{M} x_{i,j}}{M}. \tag{23}$$

is the sample mean of attribute j, and

$$s_j = \sqrt{\frac{\sum_{i=1}^{M} (x_{i,j} - \overline{x}_j)^2}{M-1}} \tag{24}$$

is the sample standard deviation.

4.3 Data Visualization with Two PCs

We perform the PCA on the normalized data. Then, we visualize the data by showing the scatter plot of the first and fourth principal components, shown in Figure 3. The color of the data points is the median house value. The two PCs show the capability for interpreting the median house value.

We also visualize the principal components one by one below. The color of dots is corresponding to the house median price. It shows that the data have limited variation on the last two principal components.

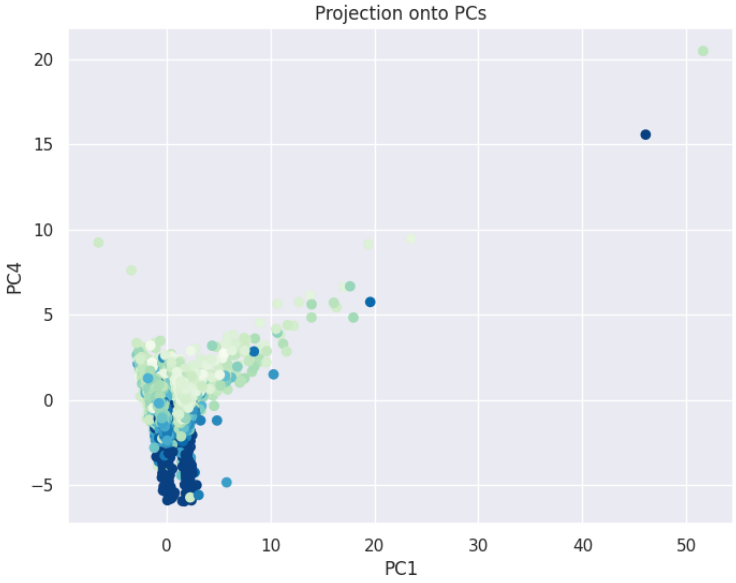


Figure 3: Visualization of the data on the first two principal components

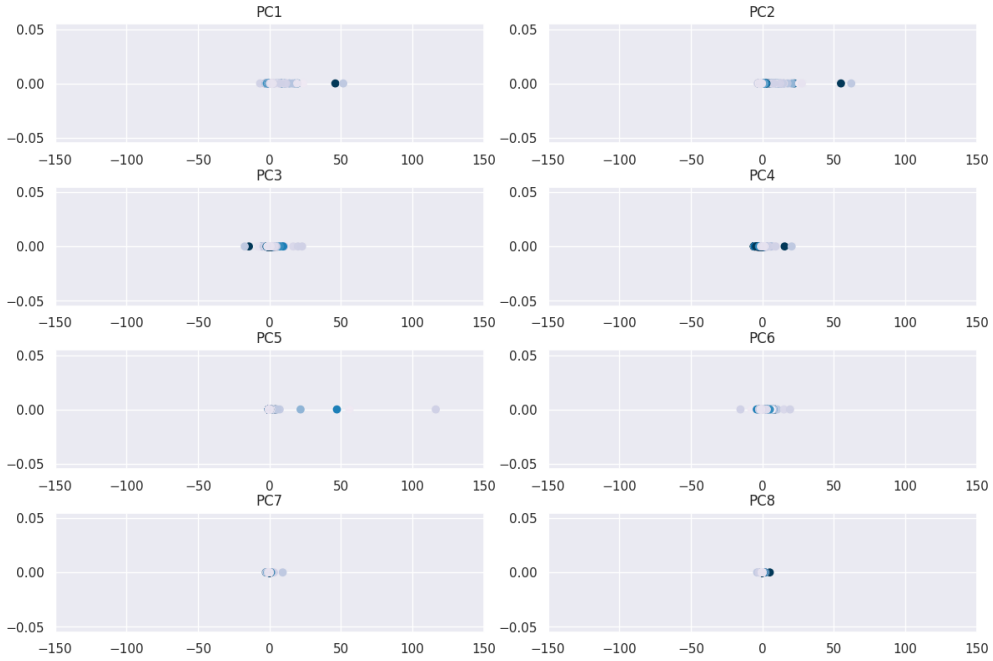


Figure 4: Visualization of the data on each principal component

4.4 Interpret Principal Components

Let

$$[\mathbf{C}_1, \ldots, \mathbf{C}_N] = \widetilde{\mathbf{X}}\mathbf{P} = [\widetilde{\mathbf{X}}\mathbf{p}_1, \ldots, \widetilde{\mathbf{X}}\mathbf{p}_N] \tag{25}$$

be the reconstruction of the normalized data. The jth principal component, $\mathbf{C}_j$, is the normalized linear combination of $\widetilde{\mathbf{X}}$'s N attributes using the jth eigen vector of $\mathbf{X}$'s covarince matrix, $\mathbf{p}_j$:

$$\mathbf{C}_j = \widetilde{\mathbf{X}}\mathbf{p}_j = [\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_N]\mathbf{p}_j = [\widetilde{\mathbf{x}}_1, \ldots, \widetilde{\mathbf{x}}_N] \begin{bmatrix} p_{1,j} \\ \vdots \\ p_{N,j} \end{bmatrix} = p_{1,j}\widetilde{\mathbf{x}}_1 + \cdots + p_{N,j}\widetilde{\mathbf{x}}_N \tag{26}$$

for $j = 1, \ldots, N$, Therefore, we can interpret $\mathbf{p}_j$'s as the loadings of $\widetilde{\mathbf{X}}$ for the linear transformation in eq. (25). Figure 5 visualizes $\mathbf{p}_j$ for $j = 1, \ldots, 8$ as rows of the heat map. The larger the value of a cell, the higher the correlation between the corresponding principal component and feature. For example, the fifth principal component and the attribute AveOccup are strongly positively correlated. The fourth principal component and the feature Medinc are strongly negatively correlated.

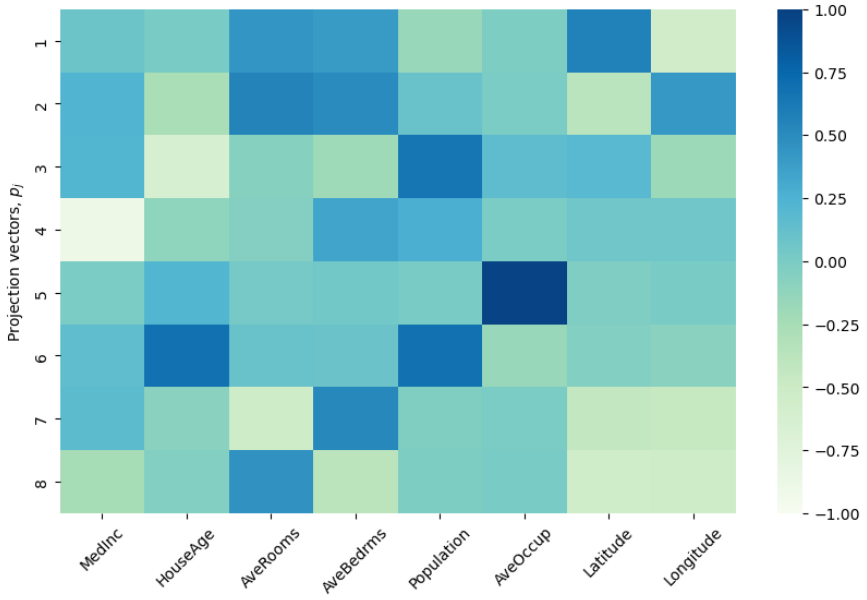


Figure 5: Interpret principal components

4.5 Explained Variance by Principal Components

In Figure 6, bars show the proportions of variance explained by individual principal components and the line shows the proportion of variances explained by the top d principal components. The figure shows the first principal component explains 25.34% of total variance. The top 5 principal components explain over 90% of total variance. Therefore, we might extract fewer principal components as new features for predicting the median value of houses. The principal components are independent because their correlation coefficients are zeros. If we build a multiple factor model using principal components for predicting the median house value, the collinearity issue is no longer an issue.

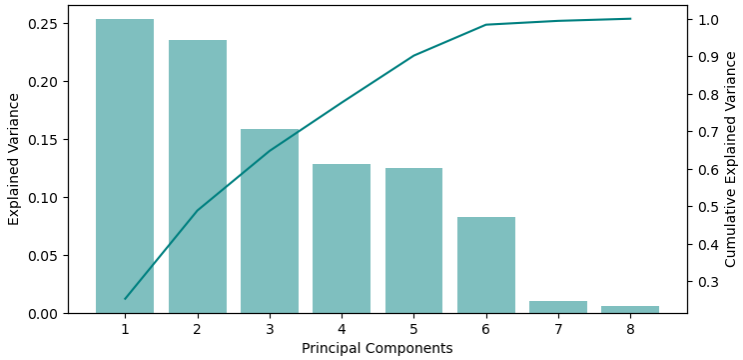
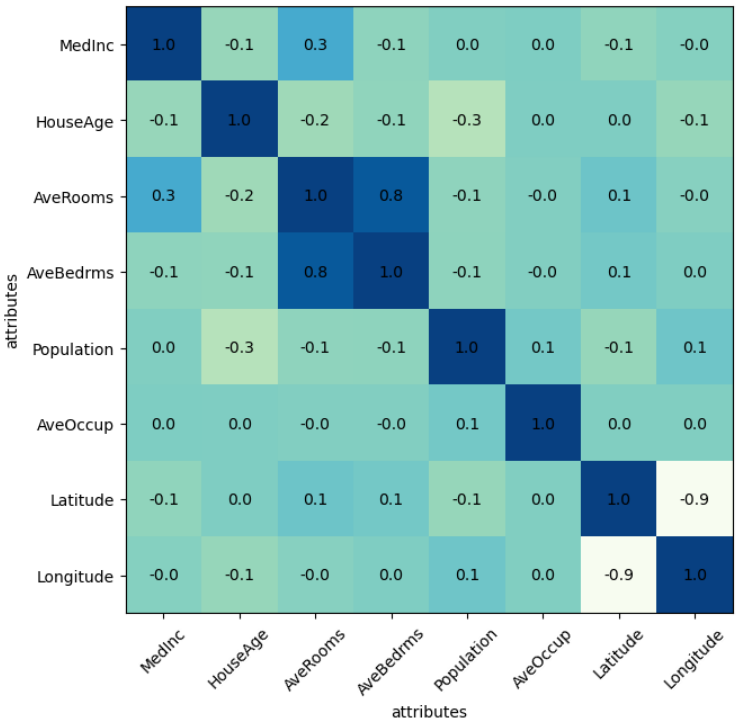


Figure 6: Explained variances by principal components

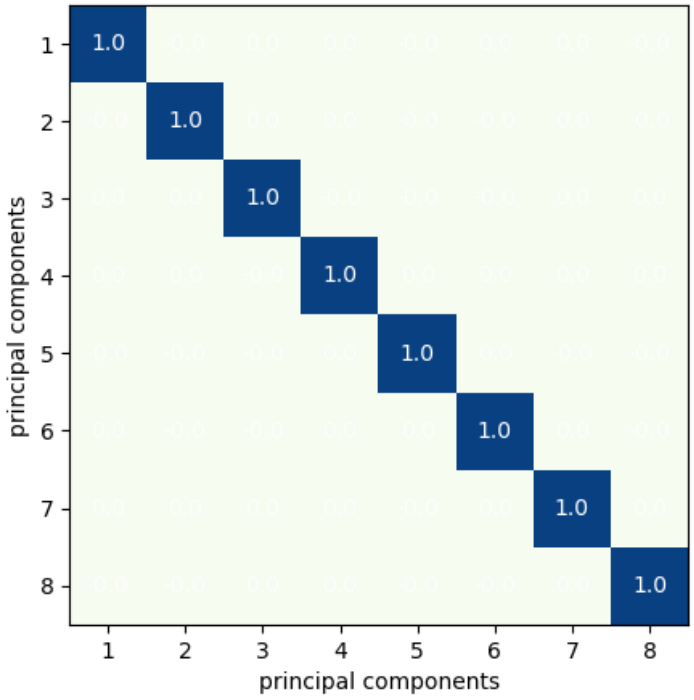
to be continued...

References

[1] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.



(a)



(b)

Figure 7: (a) correlation coefficients between attributes; (b) correlation coefficients between principal components