Big Data Management and Analytics
Assignment 9

Assignment 9-1

Consider the $X \in \mathbb{R}^{M\,x\,N}$ matrix containing six data points $X_i \in \mathbb{R}^2$.

$$X = \begin{pmatrix} 1 & 0 \\ 2 & 0 \\ 3 & 0 \\ 5 & 6 \\ 6 & 6 \\ 7 & 6 \end{pmatrix}$$

dim 1 dim 2

Conduct a PCA on the given data, i.e. project the data onto a one-dimensional space. Please state the eigenvectors, eigenvalues, covariance matrix and visualize the data before and after PCA.

Assignment 9-1

i. Center the data by substracting the mean value for each dimension:

$$\hat{\mu} = \frac{1}{N}\sum_i X_i = \binom{4}{3}$$

$$\tilde{X} = \begin{pmatrix} 1-4 & 0-3 \\ 2-4 & 0-3 \\ 3-4 & 0-3 \\ 5-4 & 6-3 \\ 6-4 & 6-3 \\ 7-4 & 6-3 \end{pmatrix} = \begin{pmatrix} -3 & -3 \\ -2 & -3 \\ -1 & -3 \\ 1 & 3 \\ 2 & 3 \\ 3 & 3 \end{pmatrix}$$

Assignment 9-1

ii. Calculate the covariance matrix $E\left[\left(x - E(X)\right) \cdot \left(X - E(X)\right)^T\right]$:

$$cov(X) \approx \hat{\Sigma} = \frac{1}{N}\tilde{X}^T\tilde{X} = \begin{pmatrix} 4,\overline{7} & 6 \\ 6 & 9 \end{pmatrix} = \begin{pmatrix} {}^{14}/_{3} & 6 \\ 6 & 9 \end{pmatrix}$$

Assignment 9-1

iii. Now compute the eigenpairs (eigenvalues, eigenvectors). Construct the eigendecomposition $\hat{\Sigma} = \hat{\hat{U}}\hat{\hat{S}}\hat{\hat{U}}^T$ with sorted eigenvalues $\widehat{\lambda_j}$ in $\hat{\hat{S}}$

Compute the eigenvalues:

$$\det\left(\hat{\Sigma} - \lambda I\right) = det\begin{pmatrix} {}^{14}/_{3} - \lambda & 6 \\ 6 & 9 - \lambda \end{pmatrix} = \left({}^{14}/_{3} - \lambda\right) \cdot (9 - \lambda) - 36$$

$$= 14 \ \cdot 3 - 36 - \frac{14+27}{3}\lambda + \lambda^2 =$$

$$\lambda^2 - \frac{41}{3}\lambda + 6 = 0$$

$$\lambda_{1,2} = \frac{{}^{41}/_{3} \mp \sqrt{\left({}^{41}/_{3}\right)^2 - 4 \cdot 6}}{2} = 13.21 \ and \ 0.45$$

Assignment 9-1

iii. Now compute the eigenpairs (eigenvalues, eigenvectors). Construct the eigendecomposition $\hat{\Sigma} = \hat{\hat{U}}\hat{\hat{S}}\hat{\hat{U}}^T$ with sorted eigenvalues $\widehat{\lambda_j}$ in $\hat{\hat{S}}$

Compute the eigenvectors:

$$\hat{\Sigma}\begin{pmatrix} x \\ y \end{pmatrix} = \lambda_{1,2}\begin{pmatrix} x \\ y \end{pmatrix}$$

$$\begin{pmatrix} 14/3 & 6 \\ 6 & 9 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \lambda_{1,2}\begin{pmatrix} x \\ y \end{pmatrix}$$

$$\overset{\lambda_1}{\Rightarrow} \quad \begin{matrix} 14/3\, x + 6y = \lambda_1 x \\ 6x + 9y = \lambda_1 y \end{matrix} \quad \Rightarrow \text{1}^{\text{st}} \text{ (normed) eigenvector:} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0{,}57 \\ 0{,}82 \end{pmatrix}$$

$$\text{Eigenvalues: } diag\left(\hat{\hat{S}}\right) = \begin{pmatrix} 13{,}21 & 0 \\ 0 & 0{,}45 \end{pmatrix}$$

$$\text{Eigenvectors: } \hat{\hat{U}} = \begin{pmatrix} 0{,}57 & 0{,}82 \\ 0{,}82 & -0{,}57 \end{pmatrix}$$

Assignment 9-1

iv. Reduce to one-dimensional space. For this purpose remove the second eigenvector and form the transformation matrix U:

$$U = \begin{pmatrix} 0{,}57 & 0 \\ 0{,}82 & 0 \end{pmatrix} \Rightarrow \begin{pmatrix} 0{,}57 \\ 0{,}82 \end{pmatrix}$$

Now transform the data with

$$Y = \tilde{X} \cdot U = \begin{pmatrix} -3 & -3 \\ -2 & -3 \\ -1 & -3 \\ 1 & 3 \\ 2 & 3 \\ 3 & 3 \end{pmatrix} \begin{pmatrix} 0{,}57 \\ 0{,}82 \end{pmatrix} = (-4{,}18 \;\; -3{,}6 \;\; -3{,}03 \;\; 3{,}03 \;\; 3{,}6 \;\; 4{,}18)^T$$

Assignment 9-1

iv. Reduce to one-dimensional space. For this purpose remove the second eigenvector and form the transformation matrix U:

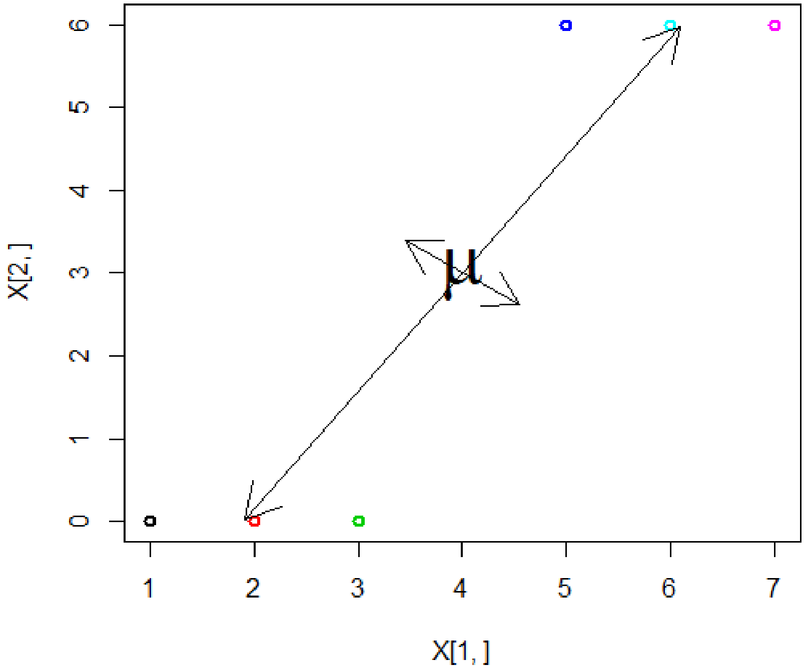
We can now try to reconstruct the original data matrix with

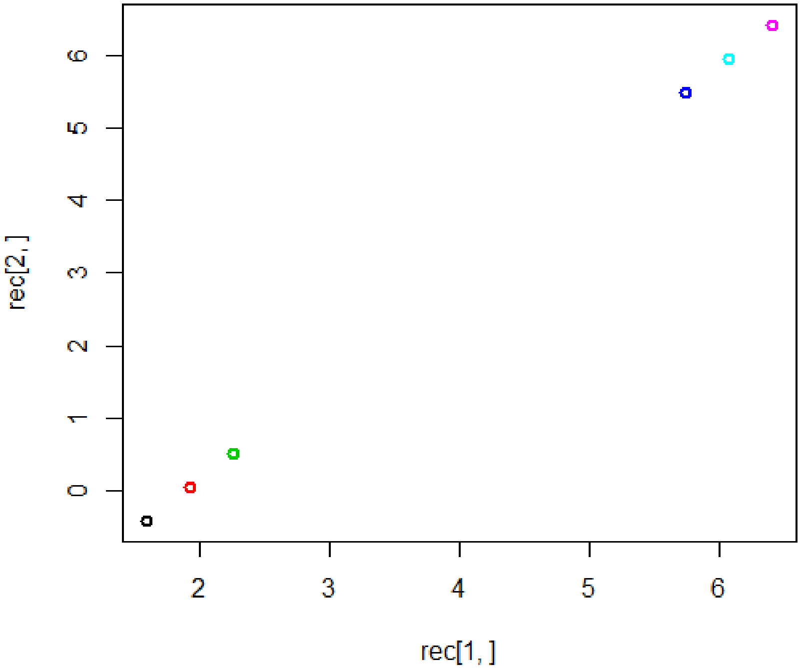
$$\hat{Z} = \mu + Y \cdot U^T = \mu + \tilde{X} \cdot U \cdot U^T$$

$$\hat{Z} = \begin{pmatrix} 1{,}6 & -0{,}42 \\ 1{,}93 & 0{,}05 \\ 2{,}26 & 0{,}52 \\ 5{,}74 & 5{,}48 \\ 6{,}07 & 5{,}95 \\ 6{,}40 & 6{,}42 \end{pmatrix}$$

Assignment 9-1

v. As we have already reduced to the one-dimensional space (here we did that by eliminiating the second principal component), the reconstruction does not imply the information of the second pc:





Assignment 9-2

- Given the matrix $M = \begin{pmatrix} 14/3 & 6 \\ 6 & 9 \end{pmatrix}$
- Determine the strongest eigenvector of M using the Power Iteration method.

```
Input: dxd data matrix M
x_0 = random unit vector
while x_i/||x_i||- x_{i-1}/||x_{i-1}||> ε do
        x_i = M^i x_0
        i=i+1
return x_i/||x_i||
```

Assignment 9-2

```
iteration:  1
x_i:  [[ 10.66666667  15.        ]]
x_i-1:  [[ 1.  1.]]
x_i_norm:  [[ 0.57952379  0.81495532]]
x_i-1_norm:  [[ 0.70710678  0.70710678]]
delta:  [[-0.12758299  0.10784854]]
----------------
iteration:  2
x_i:  [[ 139.77777778  199.        ]]
x_i-1:  [[ 10.66666667  15.        ]]
x_i_norm:  [[ 0.57478017  0.81830786]]
x_i-1_norm:  [[ 0.57952379  0.81495532]]
delta:  [[-0.00474361  0.00335254]]
----------------
iteration:  3
x_i:  [[ 1846.2962963   2629.66666667]]
x_i-1:  [[ 139.77777778  199.        ]]
x_i_norm:  [[ 0.57461679  0.8184226 ]]
x_i-1_norm:  [[ 0.57478017  0.81830786]]
delta:  [[-0.00016339  0.00011474]]
----------------
iteration:  4
x_i:  [[ 24394.04938272  34744.77777778]]
x_i-1:  [[ 1846.2962963   2629.66666667]]
x_i_norm:  [[ 0.57461117  0.81842654]]
x_i-1_norm:  [[ 0.57461679  0.8184226 ]]
delta:  [[ -5.61592869e-06   3.94293042e-06]]
----------------
convergence reached:  [[ 0.57461117  0.81842654]]
```

Assignment 9-3

Given the matrix M:

$$M = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & -1 \end{pmatrix}$$

1. Find the eigenpairs for matrix M

i. Compute: $M^T M = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$

ii. Find eigenvalues:

$$\det(M^T M - \lambda \cdot I_{2x2}) = 0$$

$$\lambda^2 - 6\lambda + 8 = (\lambda - 4)(\lambda - 2)$$

$$Eigenvalue\ \lambda_1 = 4\ \rightarrow singular\ value\ \sigma_1 = \sqrt{\lambda_1} = 2$$

$$Eigenvalue\ \lambda_2 = 2\ \rightarrow singular\ value\ \sigma_2 = \sqrt{\lambda_2} = \sqrt{2}$$

Assignment 9-3

iii. Find eigenvectors:

$$1^{st} eigenvector\ v_1: (M^T M - \lambda_1 \cdot I_{2x2})v_1 = 0 \rightarrow \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} v_1 = 0$$

$$v_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \xrightarrow{normalize} v_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$$

$$2^{nd} eigenvector\ v_2: (M^T M - \lambda_2 \cdot I_{2x2})v_2 = 0 \rightarrow \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} v_2 = 0$$

$$v_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \xrightarrow{normalize} v_2 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix}$$

iv. Eigenpairs (eigenvalue, eigenvector):

$$(\lambda_1, v_1) = \left(4, \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \right), (\lambda_2, v_2) = \left(2, \begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix} \right)$$

Assignment 9-3

2. Find the SVD for the original matrix $M = U\Sigma V^T$

From the results of (1.) we know:

$$\Sigma = \begin{pmatrix} 2 & 0 \\ 0 & \sqrt{2} \\ 0 & 0 \end{pmatrix} \, and \; V = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}$$

Assignment 9-3

- How can we find now U?
  Multiply the SVD $A = U\Sigma V^T$ with V on each side yields: $AV = U\Sigma$

$$
\begin{aligned}
U \cdot \Sigma &= (u_1\ u_2\ \dots\ u_m) \cdot \begin{pmatrix} \sigma_1 & 0 & \cdots \\ 0 & \sigma_2 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \\
&= (\sigma_1 \cdot u_1 \quad \sigma_2 \cdot u_2 \quad \dots\ \sigma_r \cdot u_r \ \ 0 \dots 0) \\
&= (A \cdot v_1 \quad A \cdot v_2 \quad \dots A \cdot v_r \ \ 0 \dots 0)
\end{aligned}
$$

Assignment 9-3

$$\Sigma = \begin{pmatrix} 2 & 0 \\ 0 & \sqrt{2} \\ 0 & 0 \end{pmatrix} \qquad V = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \qquad A = M = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & -1 \end{pmatrix}$$

- Compute

$$u_1 = \frac{1}{\sigma_1} \cdot A \cdot v_1 = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} = \frac{1}{2} \begin{pmatrix} \frac{2}{\sqrt{2}} \\ \frac{2}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$$

$$u_2 = \frac{1}{\sigma_2} \cdot A \cdot v_2 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ 0 \\ \sqrt{2} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Assignment 9-3

$$\Sigma = \begin{pmatrix} 2 & 0 \\ 0 & \sqrt{2} \\ 0 & 0 \end{pmatrix} \qquad V = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \qquad A = M = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & -1 \end{pmatrix}$$

- Having $u_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$ and $u_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ we could write now the SVD as follows:

$$M = U\Sigma V^T = \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 & * \\ \frac{1}{\sqrt{2}} & 0 & * \\ 0 & 1 & * \end{pmatrix} \cdot \begin{pmatrix} 2 & 0 \\ 0 & \sqrt{2} \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 & * \\ \frac{1}{\sqrt{2}} & 0 & * \\ 0 & 1 & * \end{pmatrix} \cdot \begin{pmatrix} \sqrt{2} & \sqrt{2} \\ 1 & -1 \\ 0 & 0 \end{pmatrix}$$

Assignment 9-3

$$M = U\Sigma V^T = \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 & * \\ \frac{1}{\sqrt{2}} & 0 & * \\ 0 & 1 & * \end{pmatrix} \cdot \begin{pmatrix} 2 & 0 \\ 0 & \sqrt{2} \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 & * \\ \frac{1}{\sqrt{2}} & 0 & * \\ 0 & 1 & * \end{pmatrix} \cdot \begin{pmatrix} \sqrt{2} & \sqrt{2} \\ 1 & -1 \\ 0 & 0 \end{pmatrix}$$

- In the last matrix multiplication →entries in the last column of U get multiplied by 0

Assignment 9-3

- How do we compute now u_3 as a third orthonormal vector?
- $\{u_1, u_2\}$ is an orthonormal basis for a plane in $\mathbb{R}^3$
  - To extend $\{u_1, u_2\}$ to an orthonormal basis for all of $\mathbb{R}^3$ we need a third vector u_3 that is normal to this plane
  - How? Compute the cross product: $u_3 = u_1 \times u_2 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{pmatrix} \times \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \\ 0 \end{pmatrix}$

- Now we have all components:

$$U = \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 & \frac{\sqrt{2}}{2} \\ \frac{1}{\sqrt{2}} & 0 & -\frac{\sqrt{2}}{2} \\ 0 & 1 & 0 \end{pmatrix} \qquad \Sigma = \begin{pmatrix} 2 & 0 \\ 0 & \sqrt{2} \\ 0 & 0 \end{pmatrix} \qquad V = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}$$

Assignment 9-3

3. Compute the one-dimensional approximation of the matrix M

The k-approximated representation is given by $M \approx U_k \Sigma_k V_k^T$. Set k=1:

$$M \approx U_k \Sigma_k V_k^T \approx \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{pmatrix} \cdot (2) \cdot \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{pmatrix} \cdot \begin{pmatrix} \frac{2}{\sqrt{2}} & \frac{2}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 0 \end{pmatrix}$$

Original matrix $M = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & -1 \end{pmatrix}$

Quiz Appendix

(a) Describe what a PCA aims for and under what circumstances it is most helpful

From the lecture slides:

- Detect hidden linear correlations
- Remove redundant and noisy features
- Interpretation and visualization
- Easier storage and processing of the data

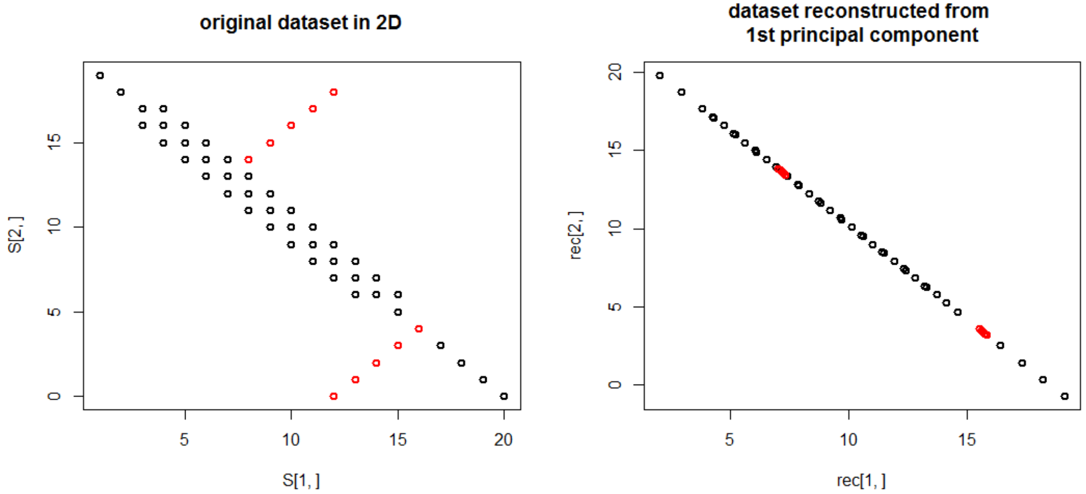
When is PCA most helpful:

- The assumption is that the observed variable can be expressed as a linear combination of the hidden variables $x = \mu + Uw + \epsilon$. If that is not the case, another heuristics should be used (e.g. LDA,RCA etc.)

Quiz Appendix

(b) Which possibly netgative consequences might arise when applying PCA to a dataset of unknown structure?

- Data which is not normed can skew the result. Therefore first norm the data!
- Loss of possibly relevant structures (see red lines within the figures)



- Solution: subspace clustering / correlation clustering

Quiz Appendix

(b) Which possibly netgative consequences might arise when applying PCA to a dataset of unknown structure?

- Further, problems with outliers may arise, as they may massively skew the PCA transformation: