Big Data Management & Analytics

EXERCISE 9 – SVD, CUR

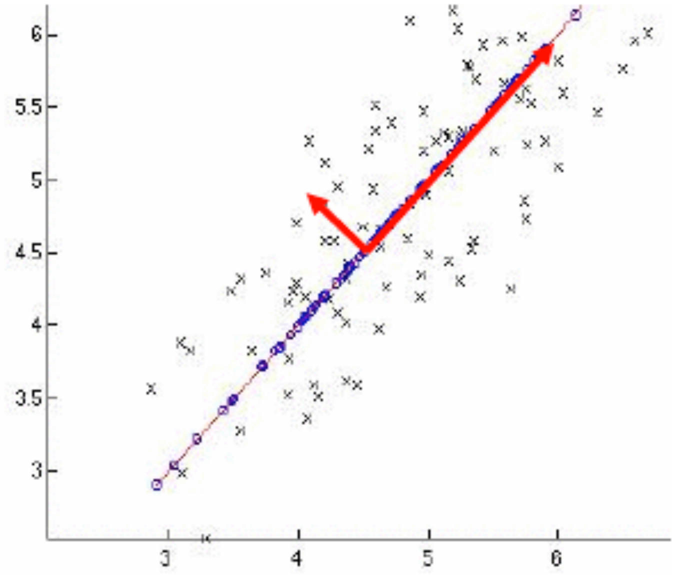
11th of January, 2016

Sabrina Friedl
LMU Munich

PCA

REVISION

PCA – Summary

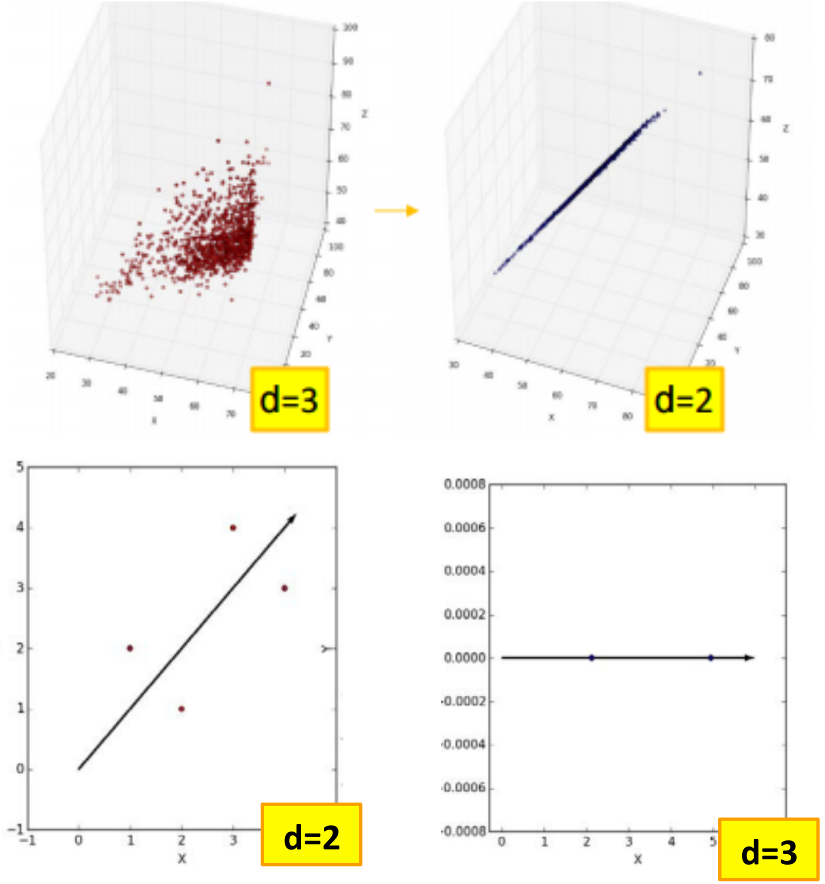


1. Center the data X : $x_i - \mu_i$
2. Calculate the covariance-matrix: $\Sigma = \frac{1}{n} X^T X$
3. Calculate the eigenvalues and eigenvectors of Σ
   - Calculate eigenvalues λ by finding the zeros of the characteristic polynomial: $\det(\Sigma - \lambda I)$
   - Calculate the eigenvectors by solving $(\Sigma - \lambda I)v = 0$
4. Select the k eigenvectors with the biggest eigenvalues and create $\mathrm{P} = (v_1, v_2, \ldots v_k)$
5. Transform the original (n x d) matrix X to a (n x k) representation: $XP = Y$

Goals of PCA

- Detect hidden correlations
- Remove redundant and noisy features
- Interpretation and visualization
- Easier storage and processing of dat

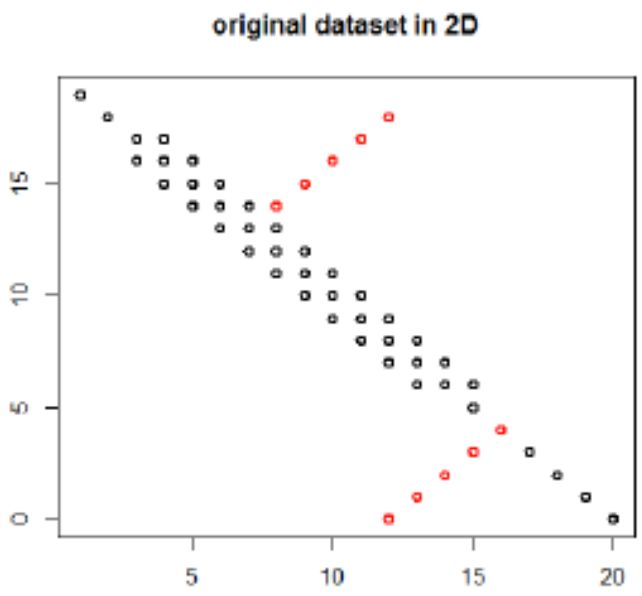
-> Most helpful when there is a linear relationship between observed and hidden variables

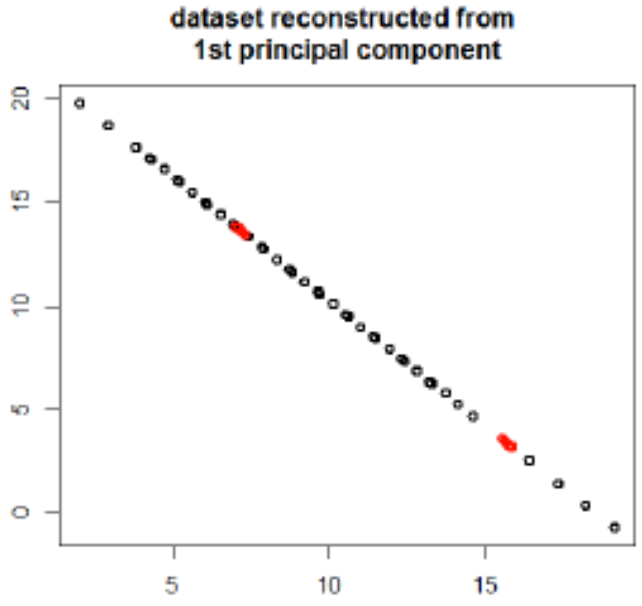


Problems with PCA

When applying PCA to a dataset of unknown structure

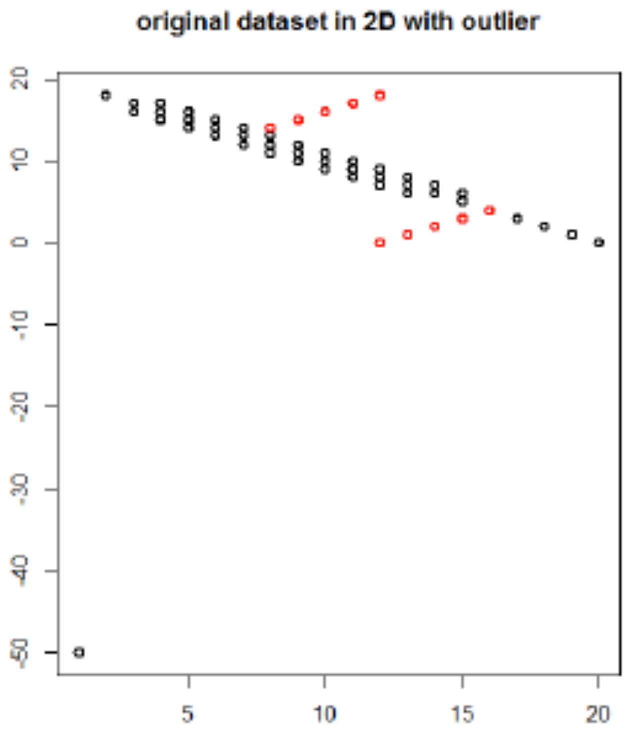
1. Unnormalized data can skew the result -> before PCA, norm the data!
2. Relevant structures might get lost

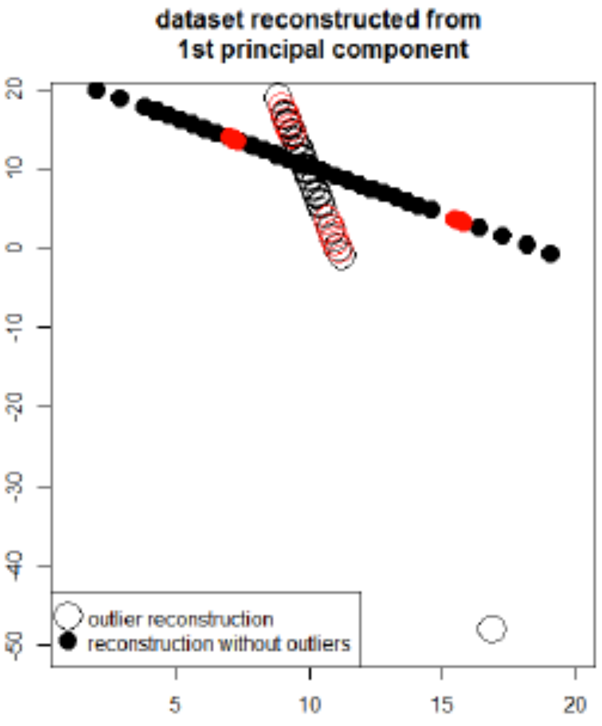




Problems with PCA

3. Outliers can skew the PCA result





Single Value Decomposition (SVD)

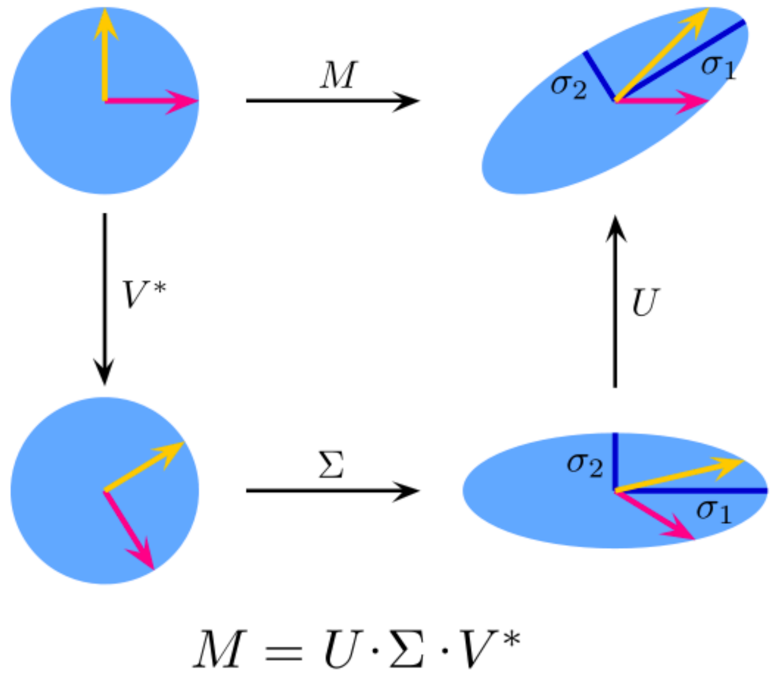
REVISION AND EXERCISE

SVD

Any matrix X can be written as $X = U \Sigma V^T$

(singular value decomposition)

- ***X*** Data matrix *(n x d)*
- ***V*** Right singular vectors: eigenvectors of $X^T X$
- ***U*** Left-singular vectors of X: eigenvectors of $X X^T$
- ***Σ*** Singular Values: square roots of eigenvalues (elements on diagonal)



https://de.wikipedia.org/wiki/Singul%C3%A4rwertzerlegung

Usage example: Image compression

SVD

Let X_{nxd} be a data matrix and let k be its rank. We can decompose X into matrices U, Σ, V as follows:

$$\underset{n \times d}{\overset{\boldsymbol{X}}{\begin{pmatrix} x_{1,1} & \dots & x_{1,d} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \dots & x_{n,d} \end{pmatrix}}} = \underset{n \times n}{\overset{\boldsymbol{U}}{\begin{pmatrix} u_{1,1} & \dots & u_{1,n} \\ \vdots & \ddots & \vdots \\ u_{n,1} & \dots & u_{n,n} \end{pmatrix}}} * \underset{n \times d}{\overset{\boldsymbol{\Sigma}}{\begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_d \end{pmatrix}}} * \underset{d \times d}{\overset{\boldsymbol{V^T}}{\begin{pmatrix} v_{1,1} & \cdots & v_{1,d} \\ \vdots & \ddots & \vdots \\ v_{d,1} & \cdots & v_{d,d} \end{pmatrix}}}$$

SVD- How to find matrices?

Remember the Eigenwertproblem:

$Av = \lambda v$ or $AT = T\Lambda$

v = eigenvector
λ = eigenvalue
T = eigenvector matrix
Λ diagonal eigenvalue matrix

For $X = U\Sigma V^T$

- Find V: $(X^T X)V = V\Sigma^2$
- Find U: $(XX^T)U = U\Sigma^2$. or use: $XV = U\Sigma$ $u_i = \frac{1}{\sigma_1} X * v_i$

SVD - Example

Given Matrix M $M = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & -1 \end{pmatrix}$ $M^T M = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}$

Eigenvalues: $det(M^T M - \lambda \cdot I_{2\times 2}) = \lambda^2 - 6\lambda + 8 = (\lambda - 4)(\lambda - 2)$

$\lambda_1 = 4 \rightarrow \text{singular value } \sigma_1 = \sqrt{\lambda_1} = 2$ $\lambda_2 = 2 \rightarrow \text{singular value } \sigma_2 = \sqrt{\lambda_2} = \sqrt{2}$

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

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

Eigenpairs

$(4, \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix})\ (2, \begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix})$

SVD - Example

Eigenvalue decomposition $X = U\Sigma V^T$

Now we already know: $\Sigma = \begin{pmatrix} 2 & 0 \\ 0 & \sqrt{2} \\ 0 & 0 \end{pmatrix}$ $V = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}$

How to find U? Multiplying the SVD $M = U\Sigma V^T$ with V on each side yields $MV = U\Sigma$

$$u_1 = \frac{1}{\sigma_1} \cdot M \cdot v_1 = \frac{\sqrt{2}}{2} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \qquad u_2 = \frac{1}{\sigma_2} \cdot M \cdot v_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

SVD - Example

Note: At this point we could write the SVD as follows:

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

How to find u_3?

$$u_3 = u_1 \times u_2 = \begin{pmatrix} \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \\ 0 \end{pmatrix}$$

u_1, u_2 and u_3 must build an orthonormal basis!

SVD - Example

One-dimensional approximation of matrix M

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

$$M \approx U_1\Sigma_1 V_1^T \approx \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{pmatrix} \cdot (2) \cdot \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 0 & 0 \end{pmatrix}$$

Recommended further reading: http://www.ams.org/samplings/feature-column/fcarc-svd

CUR

REVISION AND EXERCISE

CUR

Alternative to SVD, which better respects the structure of the data

Definition CUR : A CUR matrix decomposition is a low-rank approximation explicitly expressed in terms of a small number of *columns and rows of A*

$$\begin{pmatrix} & A & \end{pmatrix} \approx \begin{pmatrix} C \end{pmatrix} * \begin{pmatrix} U \end{pmatrix} * \begin{pmatrix} R \end{pmatrix}$$

Example

	Matrix	Alien	Star Wars	Cassablanca	Titanic
Joe	1	1	1	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	0	0	4	4
Jenny	0	0	0	5	5
Jane	0	0	0	2	2

Find CUR-decomposition of the given matrix with two rows and two columns!

Sample size $r = 2$

Steps

1. Create sample matrices C and R
2. Construct U from C and R

1a. Create sample matrix C

Sample columns for C:

Input: matrix $M \in \mathbb{R}^{m\,x\,n}$, sample size r
Output: $C \in \mathbb{R}^{m\,x\,r}$
1. **For** x = 1 : n **do**
2. P(x) = $\sum_i (m_{i,x})^2$ / $\|M\|_F^2$
3. **For** y = 1 : r **do**
4. Pick $z \in 1{:}n$ based on Prob(x)
5. C(: , y) = M(: , z) / $\sqrt{r * P(z)}$

Frobenius-Norm:

$$\|M\|_F = \sqrt{\sum_i \sum_j (m_{i,j})^2}$$

1a. Create sample matrix C

	Matrix	Alien	Star Wars	Cassablanca	Titanic
Joe	1	1	1	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	0	0	4	4
Jenny	0	0	0	5	5
Jane	0	0	0	2	2

$$\sum_i m_{i,1} = \sum_i m_{i,2} = \sum_i m_{i,3} = 1^2 + 3^2 + 4^2 + 5^2 = 51$$

$$\sum_i m_{i,4} = \sum_i m_{i,5} = 4^2 + 5^2 + 2^2 = 45$$

$FrobeniusNorm: \ \|M\|_F^2 = 243$ *= 3 * 51 + 2*45*

➔ $P(x_1) = P(x_2) = P(x_3) = \frac{51}{243} = 0.210$

➔ $P(x_4) = P(x_5) = \frac{45}{243} = 0.185$

1a. Create sample matrix C

	Matrix	Alien	Star Wars	Cassablanca	Titanic
Joe	1	1	1	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	0	0	4	4
Jenny	0	0	0	5	5
Jane	0	0	0	2	2

$$\begin{pmatrix}1\\3\\4\\5\\0\\0\\0\end{pmatrix}\frac{1}{\sqrt{r\cdot P(x_2)}} = \begin{pmatrix}1\\3\\4\\5\\0\\0\\0\end{pmatrix}\frac{1}{\sqrt{2\cdot 0.210}} = \begin{pmatrix}1.54\\4.63\\6.17\\7.72\\0\\0\\0\end{pmatrix}$$

$$C = \begin{pmatrix}1.54 & 1.54\\4.63 & 4.63\\6.17 & 6.17\\7.72 & 7.72\\0 & 0\\0 & 0\\0 & 0\end{pmatrix}$$

1b. Create sample matrix R

	Matrix	Alien	Star Wars	Cassablanca	Titanic
Joe	1	1	1	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	0	0	4	4
Jenny	0	0	0	5	5
Jane	0	0	0	2	2

$$\sum_j m_{4,j} = 5^2 + 5^2 + 5^2 = 75$$

$$\sum_j m_{5,j} = 4^2 + 4^2 = 32$$

$$FrobeniusNorm: \|M\|_F^2 = 243$$

$$P(y_4) = \frac{75}{243} = 0.309$$

$$P(y_5) = \frac{32}{243} = 0.132$$

1b. Create sample matrix C

	Matrix	Alien	Star Wars	Cassablanca	Titanic
Joe	1	1	1	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	0	0	4	4
Jenny	0	0	0	5	5
Jane	0	0	0	2	2

*Row 5 ** $\frac{1}{\sqrt{r \cdot P(y_4)}} = \frac{1}{\sqrt{2 \cdot 0.309}}$

*Row 6 ** $\frac{1}{\sqrt{r \cdot P(y_5)}} = \frac{1}{\sqrt{2 \cdot 0.132}}$

$$R = \begin{pmatrix} 6.36 & 6.36 & 6.36 & 0 & 0 \\ 0 & 0 & 0 & 7.78 & 7.78 \end{pmatrix}$$

2. Construct U from C and R

a) Create r x r matrix *W* as intersection of *C* and *R*

b) Apply SVD on $W = X\Sigma Y^T$

c) Compute Σ^+ as the pseudoinverse of Σ

d) Compute $U = Y(\Sigma^+)^2 X^T$

2. Construct U from C and R

	Matrix	Alien	Star Wars	Cassablanca	Titanic
Joe	1	1	1	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	0	0	4	4
Jenny	0	0	0	5	5
Jane	0	0	0	2	2

a) Create matrix W: $W = \begin{pmatrix} 5 & 5 \\ 0 & 0 \end{pmatrix}$

b) Apply SVD on W: $W = X\Sigma Y^T = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \sqrt{50} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}$

c) Pseudo-Inverse of Σ: $\Sigma^+ = \begin{pmatrix} \frac{1}{\sqrt{50}} & 0 \\ 0 & 0 \end{pmatrix}$

d) Calculate $U = Y(\Sigma^+)^2 X^T = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix} \cdot \begin{pmatrix} \frac{1}{50} & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{50\sqrt{2}} & 0 \\ \frac{1}{50\sqrt{2}} & 0 \end{pmatrix}$

Result of CUR decomposition

	Matrix	Alien	Star Wars	Cassablanca	Titanic
Joe	1	1	1	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	0	0	4	4
Jenny	0	0	0	5	5
Jane	0	0	0	2	2

$$C \cdot U \cdot R = \begin{pmatrix} 1.54 & 1.54 \\ 4.63 & 4.63 \\ 6.17 & 6.17 \\ 7.72 & 7.72 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{50\sqrt{2}} & 0 \\ \frac{1}{50\sqrt{2}} & 0 \end{pmatrix} \begin{pmatrix} 6.36 & 6.36 & 6.36 & 0 & 0 \\ 0 & 0 & 0 & 7.78 & 7.78 \end{pmatrix}$$