Assignment 3 - HWS24

Singular Value Decomposition

Anh Tu Duong Nguyen 2115931 - Ilias Login: anguyea
&
Anh-Nhat Nguyen 2034311 - Ilias Login: anhnnguy

November 17, 2024

Contents

1 Intuition on SVD

1.1 a)

Matrix M_1 has rank **1**, with its respective SVD components being:

$$U^T = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1 \end{bmatrix}, \quad V = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \end{bmatrix}.$$

Matrix M_2 has rank **1** with its respective SVD components being:

$$U^T = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1 \end{bmatrix}, \quad V = \begin{bmatrix} 0 & 2 & 1 & 2 & 0 \end{bmatrix}$$

Matrix M_3 has rank **1** with its respective SVD components being:

$$U^T = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1 \end{bmatrix}, \quad V = \begin{bmatrix} 0 & 1 & 1 & 1 \end{bmatrix}$$

Matrix M_4 has rank **2** with its respective SVD components being:

$$U^T = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad V = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

Matrix M_5 has rank **3** with its respective SVD components being:

$$U^T = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad V = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

Matrix M_6 has rank **2** with its respective SVD components being:

$$U^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad V = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

1.2 b)

cf. code. We realized we did matrix decompositions and not SVDs, therefore the results from 1.1 are invalid.

1.3 c)

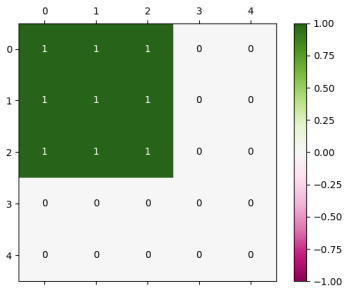


Figure 1: Rank 1 Approximation for M_1

Figure 1 displays the rank 1 approximation for the M_1 matrix that is very exact. Figure 2 displays the rank 1 to 3 approximations of M_5. Compared to the results for M_1, it can be seen that the approximations on matrix M_5 are not exact. M_1 has rank 1 and therefore a rank 1 approximation accurately reconstructs the matrix, while M_5 is of rank 3 and therefore only the rank 3 approximation can reconstruct the original matrix the best.

(a) Rank 1 Approximation for M_5

(b) Rank 3 Approximation for M_5

Figure 2: SVD approximations for rank 1 to 3

1.4 d)

It is known that M_6 is of rank 2, and therefore, the matrix should only have 2 non-zero singular values. However, when using the SVD of ***NumPy*** 5 non-zero singular vectors are returned. This could stem from the fact that we are taking roots during the computation leading to numerical instabilities.

2 SVD on Weather Data

2.1 a)

Normalization is necessary because the climate dataset contains features with different scales, and SVD is sensitive to such disparities. This approach assumes that all features are equally significant.

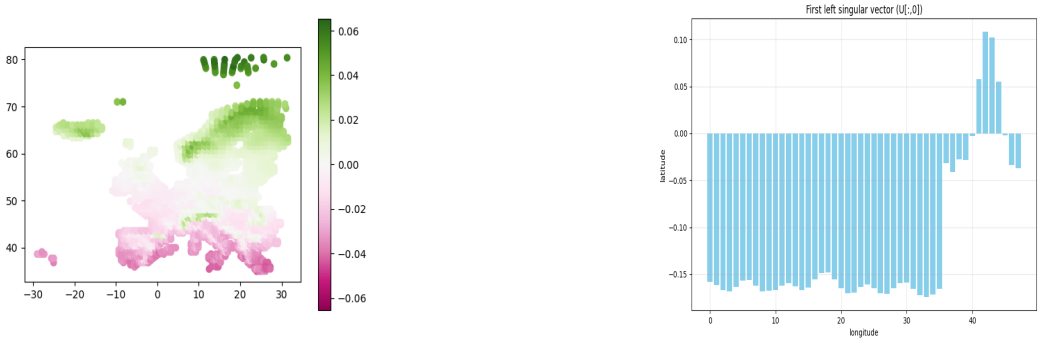
2.2 b)

Rank: 48. Detailed implementation cf code.

2.3 c)

Figure 3 reveals that the first right singular vector 3b is primarily influenced by temperature features (1–36), while its association with rainfall features is minimal. On the other hand, the first left singular vector assigns higher values to data points in the northern

regions and lower values to those in the south. This indicates an inverse relationship with temperature: regions with smaller vector values tend to experience warmer conditions, whereas larger values correspond to cooler temperatures. This relationship holds true for minimum, maximum, and average temperature metrics.

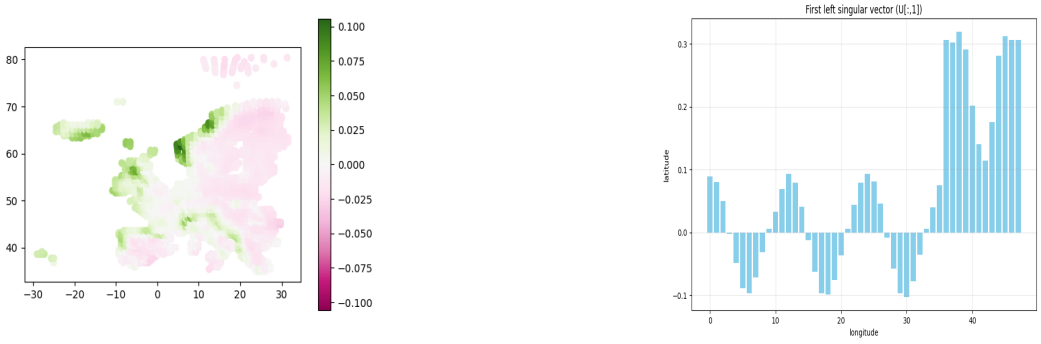


(a) First left

(b) First right

Figure 3: The first singular vector

In Figure 4, the second left singular vector assigns higher values to data points in coastal and mountainous regions, reflecting their distinct climatic conditions. This vector strongly loads on rainfall features, with temperature features showing an inverse seasonal relationship: higher values in autumn and winter (cooler months) and lower values in spring and summer (warmer months). This suggests it captures average rainfall patterns. Subfigure 4a shows elevated values in coastal areas due to high evaporation near water bodies and in mountainous regions due to orographic lifting, where moist air cools and condenses as it rises over terrain. This aligns with the inverse temperature relationship, as precipitation is greater in cooler months than in warmer ones.



(a) Second left

(b) Second right

Figure 4: The second singular vector

As shown in Figure 5, the third left singular vector highlights data points associated with significant temperature fluctuations. These include extremely low minimum temperatures and exceptionally high maximum temperatures, as reflected by the corresponding right singular vector. Furthermore, substantial differences in precipitation are evident between the summer and winter seasons. This suggests that the third left singular vector captures annual temperature variability. Data points with high loadings indicate pronounced seasonal changes, whereas lower loadings represent regions with more consistent weather patterns throughout the year.

(a) Third left (b) Third right

Figure 5: The third singular vector

Figure 6 appears to reflect elevation when compared to a topographic map of Europe. Subfigure 6a highlights that regions with high values in the fourth left singular vector correspond to areas of significant altitude, such as the Alps and the Pyrenees, while regions with low values are closer to sea level. Combined with the patterns observed in subfigure 6b, this suggests that higher-altitude regions typically exhibit characteristics of a continental climate, including extreme temperatures—low minimum and high maximum—and substantial rainfall during the summer months.

(a) Fourth left (b) Fourth right

Figure 6: The fourth singular vector

The fifth left singular vector does not provide any meaningful interpretation and is omitted here. The reason for this will become evident later when determining the optimal rank k for the truncated SVD.

2.4 d)

The first and second left singular vectors can be interpreted as representing (inverse) average temperature and rainfall, respectively. Figure 7a illustrates significant temperature variation along the x-axis, transitioning from north to south (green to red). However, the data is less distinctly separated by rainfall when analyzed in the same direction. An interesting pattern emerges for points located far to the north, which experience extremely low temperatures and minimal rainfall—characteristic of polar climates, known for being cold and dry. In contrast, Figure 7b shows that moving from east to west along the x-axis does not reveal clear separations for either temperature or rainfall. Nonetheless, a subtle distinction in rainfall is visible, which aligns with the expectation of higher precipitation along western coastal regions compared to the east.

(a) (b)

Figure 7: Rainfall vs Temperature

Since the other columns of U did not provide any significant or clear insights, their plots have been omitted. For additional information, please consult the Jupyter Notebook.

2.5 e)

(i) Based on the Guttman-Kaiser criterion, the optimal SVD rank (k) is determined to be 37.
(ii) As illustrated in Figure 9, 90% of the squared Frobenius norm is achieved at $k = 3$.

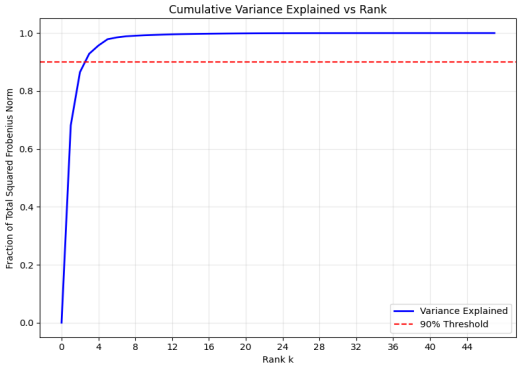


Figure 8: 90 percent of squared Frobenius norm

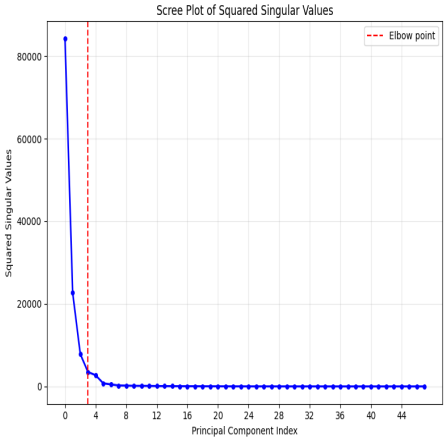


Figure 9: Scree Plot of SSV

(iii) Using Cattell's Scree test, the optimal size is determined to be $k = 4$.
(iv) According to the entropy-based method and Figure 10, the optimal rank is $k = 1$.

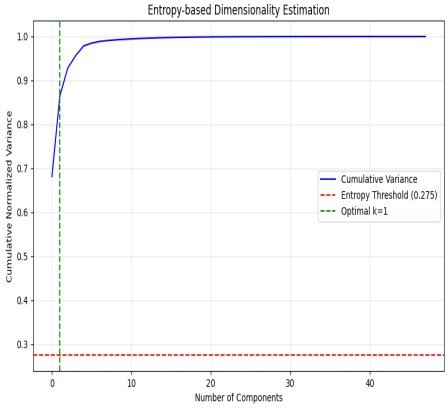


Figure 10: Entropy-based plot

(v) As shown in Figure 12, the optimal rank is $k = 8$.

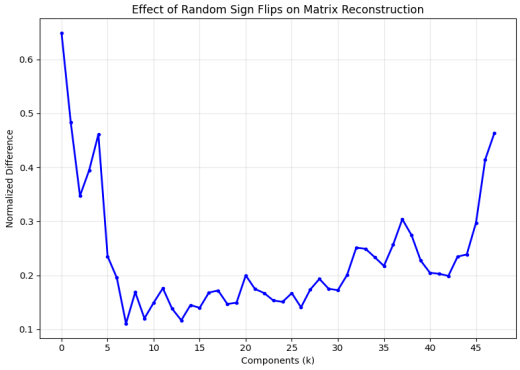


Figure 11: Random Sign Flips

2.6 f)

At first glance, using the full SVD with no noise ($\epsilon = 0$) naturally results in a perfect reconstruction, yielding an RMSE of zero. A less intuitive result, however, is that introducing noise to the full reconstruction produces a higher RMSE compared to a truncated SVD with $k \in \{1, 2, 5, 10\}$. This observation aligns with the role of SVD in denoising data, where reduced-rank approximations can eliminate smaller singular values associated with noise. For $k = 1$, the RMSE remains unaffected by the noise level (ϵ) since only the dominant singular value is retained, effectively filtering out noise entirely. In contrast, when $k = 2$, the reconstruction starts to degrade for $\epsilon > 0.5$. Interestingly, at higher noise levels (e.g., $\epsilon = 2$), the $k = 2$ approximation outperforms other configurations, achieving the lowest RMSE. This highlights the balance between rank selection and noise handling in SVD-based reconstructions.

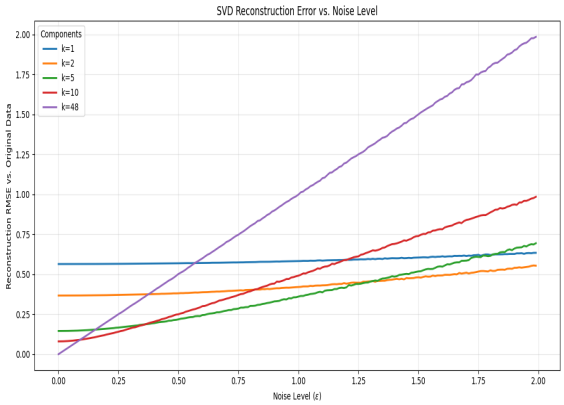


Figure 12: Impact of Noise Levels on RMSE for Truncated SVD Reconstructions

3 SVD and Clustering

3.1 a)

Figure 13 shows K-Means Clustering for k=5 clusters interpretable as different climate regions in the EU.

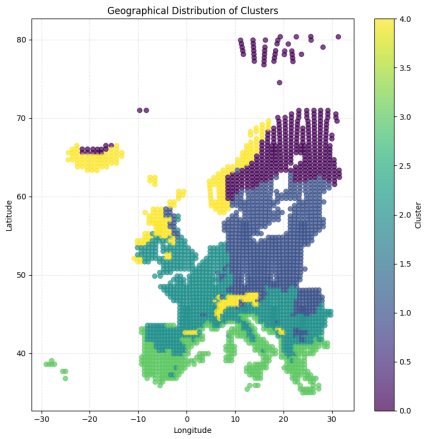


Figure 13: Clusters based on K-Means

3.2 b)

Figure 14 shows that the cluster borders have smooth edges, meaning that the clusters are well separated. and do not overlap. An exception is the yellow cluster. A wider variation is observed and it can be argued that we have a separate cluster in that region.

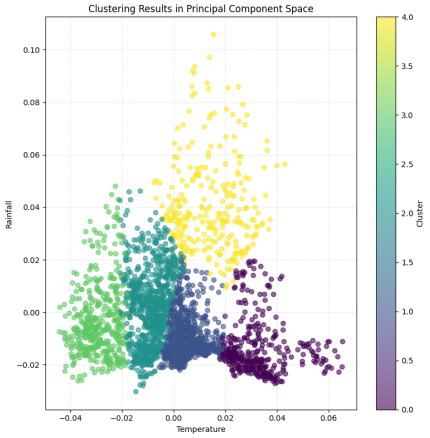


Figure 14: 1st and 2nd singular vector colour coded by cluster

3.3 c)

Computing the PCA scores of a matrix $\mathbf{X}$ for the first $\boldsymbol{k}$ principal components requires multiplying the first $\boldsymbol{k}$ left singular vectors and singular values from the SVD of $\mathbf{X}$. This approach works since the data was normalized before the SVD computation. For $\mathbf{k}$=1, there are significant differences to the original clustering, especially in the northeastern region. For the other two PCA scores, it is not possible to discover any noticeable differences compared to the original clustering. PCA is a form of dimensionality reduction, where the largest directions of variations in the data are inspected. This means that the data is reduced to the k-th dimension, with k $\in$ 1, 2, 3.It was discussed that using the truncated SVD is equivalent to noise reduction and therefore K-Means clustering on truncated data shows similar performance when compared to the K-Means clustering on the original data.

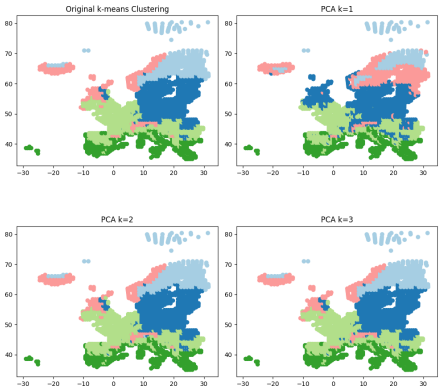


Figure 15: k-means and PCA based clusters for k=1 to 3

Ehrenwörtliche Erklärung

Ich versichere, dass ich die beiliegende Bachelor-, Master-, Seminar-, oder Projektarbeit ohne Hilfe Dritter und ohne Benutzung anderer als der angegebenen Quellen und in der untenstehenden Tabelle angegebenen Hilfsmittel angefertigt und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen. Ich bin mir bewusst, dass eine falsche Erklärung rechtliche Folgen haben wird.

Declaration of Used AI Tools

Tool	Purpose	Where?	Useful?
ChatGPT	Rephrasing	Throughout	+
DeepL	Translation	Throughout	+
ResearchGPT	Summarization of related work	Sec. **??**	-
Dall-E	Image generation	Figs. 2, 3	++
GPT-4	Code generation	functions.py	+
ChatGPT	Related work hallucination	Most of bibliography	++

Unterschrift
Mannheim, den XX. XXXX 2024