Assignment Report

Assignment 3: Singular Value Decomposition

Artem Bisliouk (abisliou), Elizaveta Nosova (enosova)

November 17, 2024

1. Introduction to SVD

a) Manual estimation of matrix rank and SVD

Detailed derivations of the results represented in this section are provided in Appendix A.

Matrix $\mathbf{M_1}$

- **Rank:** 1
- **Singular Values:** $\sigma_1 = 3$
- **Left Singular Vectors** (U)**:** $\mathbf{u}_1 = \frac{1}{\sqrt{3}}[1, 1, 1, 0, 0]^\top$.
- **Right Singular Vectors** (V)**:** $\mathbf{v}_1 = \frac{1}{\sqrt{3}}[1, 1, 1, 0, 0]^\top$.

Matrix $\mathbf{M_2}$

- **Rank:** 1
- **Singular Values:** $\sigma_1 = 3\sqrt{3}$
- **Left Singular Vectors** (U)**:** $\mathbf{u}_1 = \frac{1}{\sqrt{3}}[0, 1, 1, 1, 0]^\top$.
- **Right Singular Vectors** (V)**:** $\mathbf{v}_1 = \frac{1}{3}[0, 2, 1, 2, 0]^\top$.

Matrix M_3

- **Rank:** 1
- **Singular Values:** $2\sqrt{3}$
- **Left Singular Vectors (U):** $\mathbf{u}_1 = \frac{1}{2}[0,1,1,1,1]^\top$.
- **Right Singular Vectors (V):** $\mathbf{v}_1 = \frac{1}{\sqrt{3}}[0,1,1,1]^\top$.

Matrix M_4

- **Rank:** 2
- **Singular Values:** $\sigma_1 = 3, \quad \sigma_2 = 2$
- **Left Singular Vectors (U):** $\mathbf{u}_1 = \frac{1}{\sqrt{3}}[1,1,1,0,0]^\top, \quad \mathbf{u}_2 = \frac{1}{\sqrt{2}}[0,0,0,1,1]^\top$.
- **Right Singular Vectors (V):** $\mathbf{v}_1 = \frac{1}{\sqrt{3}}[1,1,1,0,0]^\top, \quad \mathbf{v}_2 = \frac{1}{\sqrt{2}}[0,0,0,1,1]^\top$.

Matrix M_5

- **Rank:** 3
- **Singular Values:** FAILED
- **Left Singular Vectors (U):** FAILED
- **Right Singular Vectors (V):** FAILED

Matrix M_6

- **Rank:** 2
- **Singular Values:** FAILED
- **Left Singular Vectors (U):** FAILED
- **Right Singular Vectors (V):** FAILED

b) SVD computation with NumPy

Comparison for M_1, M_2, M_3, M_4. The manual computations for M_1, M_2, M_3, M_4 were correct in terms of rank, singular values, and the dominant singular vectors. The discrepancies in the signs of singular vectors between the manual and NumPy computations are valid, as singular vectors are unique up to sign. The extremely small singular values obtained from NumPy, which are close to zero, result from numerical precision limitations inherent in floating-point arithmetic. These values do not contradict the manual results, as they effectively represent zero, but they lead to the non-zero values in the corresponding left and right singular vectors.

Comparison for M_5 and M_6. The manual computations for M_5 and M_6 correctly identified their ranks as 3 and 2, respectively, which align with the NumPy results. For both matrices, the inability to compute the left and right singular vectors manually was acknowledged due to the non-orthogonality of $M_5 M_5^T$, $M_5^T M_5$, $M_6 M_6^T$, and $M_6^T M_6$. NumPy confirmed this complexity by providing precise singular vectors. The inferred singular values for M_6 matched well with NumPy's computed values ($\sigma_1 = 4.8284$, $\sigma_2 = 0.8284$). For M_5, the inferred alignment of singular values with row groups was consistent, but exact values ($\sigma_1 = 3.5616$, $\sigma_2 = 2.0000$, $\sigma_3 = 0.5616$) could only be determined computationally.

c) Best rank-1 approximation

The rank-1 approximations across M_1 to M_6, as shown in Appendix B, reveal consistent patterns and intuitive insights into the dominant structures of these matrices. For M_1 (Figure 3), M_2 (Figure 4), and M_3 (Figure 5), the approximations effectively capture their repeated row groups ($[1, 1, 1, 0, 0]$, $[0, 2, 1, 2, 0]$, and $[0, 1, 1, 1]$, respectively), emphasizing their low-rank nature and uniform structure. These results are highly intuitive, as these matrices have a single dominant pattern, and their approximations faithfully retain the core features while minimizing the Frobenius norm of the difference.

In contrast, M_4 (Figure 6) demonstrates the ability of the rank-1 approximation to focus on the most influential row group ($[1, 1, 1, 0, 0]$) while effectively ignoring the smaller secondary group. This highlights how the method prioritizes the largest contribution to the matrix's structure, making it an intuitive choice for matrices with distinct but unequal row groups.

For M_5 (Figure 7), the approximation balances multiple row groups, resulting in a more averaged representation. Unlike M_1-M_4, where a single dominant pattern is clear, the interplay of distinct groups in M_5 introduces complexity, leading to a less intuitive but mathematically optimal approximation. Similarly, M_6 (Figure 8) smooths out its irregularities, particularly the deviation in the third row, while maintaining the homogeneity of its overall structure. This highlights how rank-1 approximations can adapt to minor deviations while preserving the matrix's dominant patterns.

Overall, the rank-1 approximations consistently capture the essence of the matrices, prioritizing dominant patterns and structural simplicity. While matrices with a single clear row group (M_1-M_4) produce more intuitive approximations, those with mixed structures (M_5 and M_6) require a trade-off, resulting in a less visually distinct but still optimal representation.

d) Singular values of matrix M6

The rank of M_6, determined using the `np.linalg.matrix_rank` function, is 2, which matches the number of linearly independent rows in the matrix. This is consistent with the theoretical understanding that M_6 contains two independent row patterns: the dominant uniform rows and the deviation in the third row.

NumPy's `svd` function reports five singular values: $\sigma_1 = 4.83$, $\sigma_2 = 0.83$, and three additional values very close to zero ($\sigma_3 = 9.95 \times 10^{-17}$, $\sigma_4 = 2.19 \times 10^{-17}$, $\sigma_5 = 5.32 \times 10^{-50}$). These extremely small singular values potentially arise due to numerical precision limitations in floating-point arithmetic and are effectively treated as zero.

2. The SVD on Weather Data

a) Data Normalization

The climate matrix was normalized to z-scores, an essential step to account for differing units of measurement (e.g., degrees for temperature and millimeters for rainfall) and to prevent attributes with larger scales, like rainfall, from dominating. The assumption of equal importance for all attributes is reasonable, as minimum, maximum, and average temperatures, along with rainfall for each month, collectively capture seasonal variations, extremes, and climate dynamics.

As to the normal distribution assumption, real-world climate data is not necessarily normally distributed due to natural variations, as seen in plots in Figure 9 in Appendix C, which reveal some skewness in rainfall features. However, the distributions are approximately normal, allowing us to reasonably accept the assumption for normalization.

b) SVD and Rank of Normalized Data

SVD was computed with NumPy in the section 2b of the Jupiter Notebook, fragments of the matrices S and Vt are presented in Figures 11 and 10 in Appendix D. The rank of the normalized data is 48, which reveals lack of linear dependency between the features.

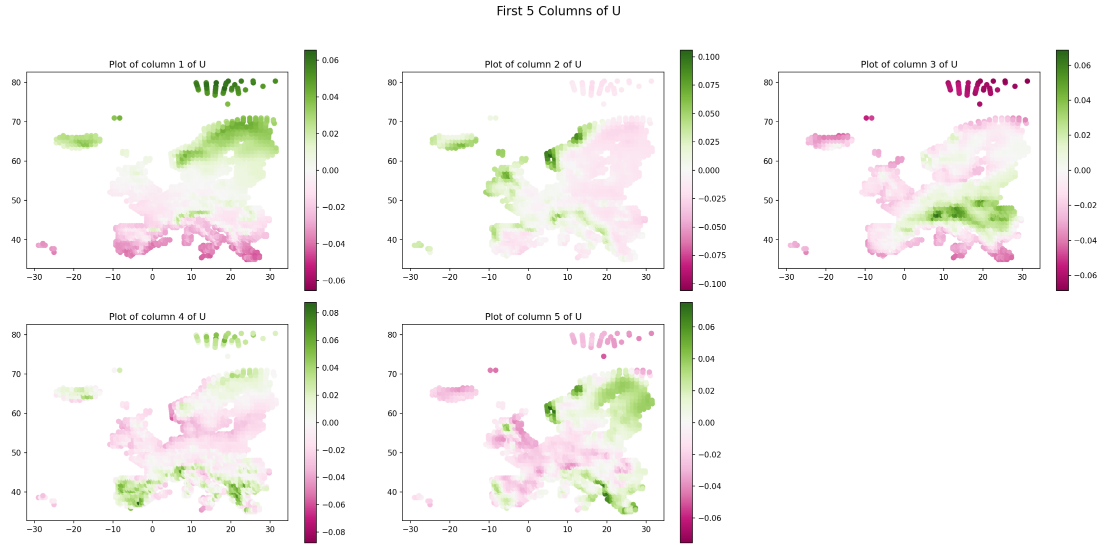
c) Interpretation of the First 5 Left Singular Vectors

In Figure 1 the first 5 columns of the U matrix (left singular vectors) are plotted, with longitude and latitude as coordinates and entries as color values. Values of the corresponding rows of Vt need to be taken into account for interpretation and can be found in Figure 10 in Appendix D.

The first column of U highlights northern regions with cooler yearly temperatures, drier winters, and wetter summers (positive values) and southern regions with warmer temperatures, wetter winters, and drier summers (negative values). Vt weights suggest this represents a tendency toward low temperatures (around -0.16 for all values projected from temperature features) and rainy summers (0.06-0.11 for rain in June-September).

The second column of U emphasizes rainfall, especially in winter months, with positive values concentrated in western and northern coastal regions and negative values in eastern and southern areas. Vt weights confirm rainfall's dominance with weight contributions of 0.28-0.31 from September to April and 0.11-0.2 in the remaining months.

The third column of U distinguishes areas with hot, wet summers and cold, dry winters (positive values) from coastal areas, especially western shores, with moderate climates (negative values). Vt highlights rainfall in May–August (e.g. 0.41 for May and 0.46 for June) and summer maximum temperatures (in range 0.16-0.22).


First 5 Columns of U

Plot of column 1 of U

Plot of column 2 of U

Plot of column 3 of U

Plot of column 4 of U

Plot of column 5 of U

Figure 1: Plot of 5 first columns of U component matrix of normalized data

The fourth column of U identifies regions with high temperature amplitudes within the same month, rainy springs, and dry summers, such as Southern Europe, Northern Scandinavia, and islands in the Arctic Ocean (positive values), while Central Europe shows more stable monthly temperatures (negative values). Vt indicates strong variation in temperatures (weights from -0.1 to -0.25 for minimal temperatures and from 0.1 to 0.22 for maximal ones) and rainfall (negative in summer, positive in spring).

The fifth column of U represents regions with more difference between summer and winter temperatures and precipitations, with positive values in the northeast and southeast and negatives in Western Europe, Iceland and northern islands. Vt highlights seasonal rainfall and temperature variations, with large positive values (0.18-0.23) for temperatures in summer and rainfall in winter and large negative values (from -0.14 to -0.24) for the opposite pattern.

d) Interpretation of Left Singular Vectors Based on North-South and East-West Distinction

In the Figure 12 and Figure 13 in Appendix E we present the plots of the first column of U against the other left singular vectors up to the fifth, with data points colored by their North-South and East-West location respectively. The first singular vector U1 defines one small and two larger distinct clusters: the islands in the Arctic Ocean (dark green), Northern Europe, and Southern Europe. The influence of the other singular vectors (on the y-axis of the plots) slightly adjusts the clusters but doesn't disrupt the core grouping, reflecting the dominance of the first singular value (s1 = 290 against s2=151 and lower for upcoming singular values). The East-West location does not significantly impact the clustering in U1, as points from both directions overlap.

Method	Selected rank
Guttman–Kaiser criterion	37
90% of squared Frobenius norm	3
Scree test	6
Entropy-based method	1
Random flipping of signs	7

Table 1: Results of different rank selection methods for truncated SVD

U2 captures more of the East-West distribution, as can be seen in Figure 14 in Appendix E. When plotted against U3, Eastern points tend to have lower U2 values compared to Western ones, though the difference in U3 is less pronounced: any data point can have low and moderate values, but only western and central points show large values. A small cluster of Eastern and Northern points shows both U2 and U3 values near their minimums, but overall, no strong clustering pattern is observed.

e) Rank Selection Methods

The suggested ranks for truncated SVD from various methods are summarized in Table 1, ranging from k=1 (entropy-based method) to k=37 (Guttman-Kaiser criterion).

For the Scree test (Figure 15 in Appendix F), we selected k=6 as the elbow point, where singular values begin to flatten: although the change between k=4 and k=5 is minimal, the drop between k=5 and k=6 remains noticeable. This choice aligns with the result of a knee locator algorithm, which identifies the maximum curvature at k=6.

For the random flipping of signs (Figure 16), the minimum relative change (0.111) in spectral norms of the original and randomly perturbed residual matrices occurs at k=7, suggesting this rank as optimal.

Given the variability across methods, we avoid blindly relying on any single suggestion. Based on the S matrix in Figure 11 and previous analysis of the first five singular vectors in Sections 2 and 2, we recommend exploring k between 3 and 6. This range balances capturing essential structure while avoiding noise, ensuring a compact yet informative representation.

f) SVD and Noise

Figure 2 shows the impact of rank k and noise level on RMSE for rank-k truncated SVD reconstruction. For k=1, RMSE is highest at low noise levels and remains nearly constant as the noise level increases. This occurs because the first singular vector captures only dominant data features, but even if they are not sufficient to capture all the subtleties, it is unaffected by noise.

For higher k, RMSE decreases at low noise levels, reflecting improved reconstruction accuracy. However, as the noise level increases, RMSE grows faster for larger k, as more

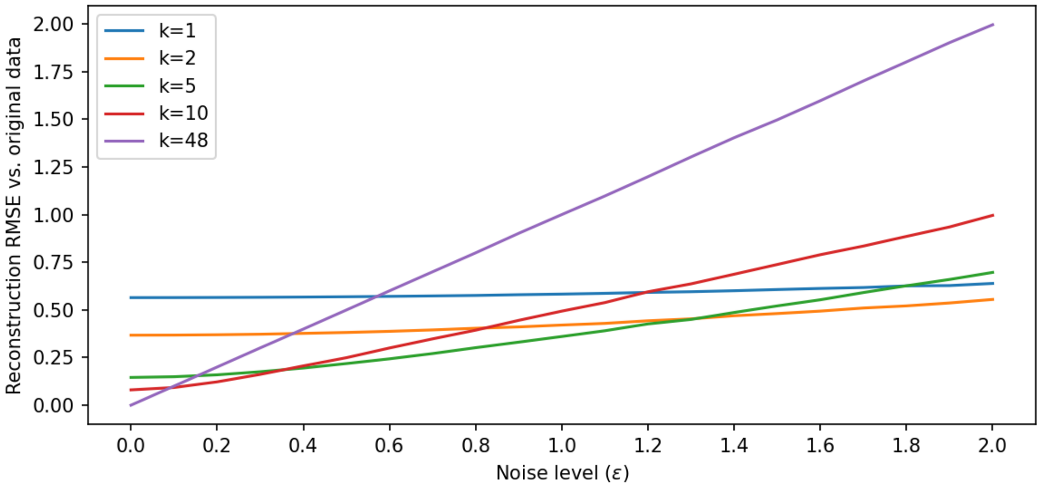


Figure 2: Impact of Noise on RMSE for Rank-l Approximations

singular vectors begin to incorporate noise. Full-rank reconstruction (k=48) perfectly approximates clean data but is most sensitive to noise.

In practical settings, the level noise is often unknown. We recommend to find a tradeoff depending on expectation of noise level: smaller k (e.g., k=2 for this case) minimizes noise sensitivity but sacrifices detail, while moderate k (e.g., k=5) balances reconstruction quality and robustness to noise.

3. SVD and Clustering

a) Cluster Interpretation

Figure 17 in Appendix G shows the clusters from K-Means with K=5, which likely correspond to different climatic zones based on temperature and rainfall patterns. The result can be interpreted as follows: the southern cluster (dark blue) might have higher temperatures, while the northern clusters (dark green and pink) are colder. The light-blue and light-green clusters in central regions might show intermediate temperatures and, most probably, differ in rainfall amount. In particular, the light-blue and pink clusters in Eastern Europe and Scandinavia are likely to have low rainfall, with the pink cluster experiencing colder, almost subarctic conditions. Rainfall generally can be increasing from light-blue to light-green clusters, peaking further in the dark-green cluster with an oceanic climate.

b) Alternative Visualization of Clusters using Singular Vector Space

As shown in Figure 18 in Appendix G, the clusters are close, with some overlap at the borders, but still distinguishable. The first left singular vector, which reflects the North-

South temperature trend, shows the highest values in the pink cluster and the lowest in the dark-blue one. This aligns with our interpretations of the vector in Section 2 and of the clusters in Section 3. The second left singular vector distinguishes East-West distribution, with dark-green and light-blue clusters showing different rainfall patterns, again consistent with observations in Section 2 and assumptions in Section 3. A tight group of pink points in the bottom right suggests it may form a separate cluster, while the dark-green cluster appears dispersed, with some points potentially being outliers.

To verify, we reran the K-Means with K=7 (Figure 19 in Appendix G) and found the bottom-right points form their own cluster. The dark-green cluster remained vertically dispersed, while light-blue and light-green clusters split into three, with some dark-green border points reassigned. Further increasing K might reveal more detailed clustering, but this is outside the scope of the current report.

c) Principal Component Analysis

In Figure 20 in Appendix G we can compare the clustering of the original data and PCA with 1, 2 and 3 principal components respectively. The clusters between the original data and the PCA-reduced data with k=1 have some difference, most noticeably in dark-green and pink clusters. This suggests that the reduced dimensions capture only the most significant features of the data resulting in coarser clustering, possibly omitting details critical to finer cluster distinctions.

For k=2,3, more features are retained, so clustering results are very close to the original data, almost unchanged. This implies that most clustering-relevant information is in the first two or three principal components.

Appendix

A. Manual SVD derivation

Manual SVD derivations in Task 1a are defined as follows:

Matrix $\mathbf{M_1}$

$$M_1 = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

- **Rank:** The first three rows of M_1 are identical: $[1, 1, 1, 0, 0]$, forming one linearly independent direction. All other rows are either identical or zero. Since there is only one unique non-zero row direction, the matrix rank is 1.

- **Singular Values:** Singular values correspond to the square roots of the eigenvalues of $M_1 M_1^T$. The product is:

$$M_1 M_1^T = \begin{bmatrix} 3 & 3 & 3 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

  Taking into account the property that the trace provides the sum of eigenvalues and for rank-1 matrices like $M_1 M_1^T$ only one eigenvalue is non-zero, the trace itself is the non-zero eigenvalue. The trace is:

$$\text{Trace}(M_1 M_1^T) = 3 + 3 + 3 + 0 + 0 = 9.$$

  Hence, the singular value is:

$$\sigma_1 = \sqrt{9} = 3.$$

- **Left Singular Vector (U):** The left singular vectors are derived from the eigenvectors of $M_1 M_1^T$. As shown above, all non-zero rows are proportional to $[1, 1, 1, 0, 0]$. Normalizing this eigenvector gives:

$$\mathbf{u}_1 = \frac{1}{\sqrt{3}}[1, 1, 1, 0, 0]^\top.$$

  This vector represents the dominant direction of the row space, as all non-zero rows of M_1 are multiples of $[1, 1, 1, 0, 0]$.

- **Right Singular Vector (V):** The right singular vectors correspond to the eigenvectors of $M_1^T M_1$. The product is:

$$M_1^T M_1 = \begin{bmatrix} 3 & 3 & 3 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

  The dominant eigenvector is proportional to $[1, 1, 1, 0, 0]$, as all columns of M_1 contribute equally to this vector. Normalizing this gives:

$$\mathbf{v}_1 = \frac{1}{\sqrt{3}}[1, 1, 1, 0, 0]^\top.$$

  This vector reflects how the columns contribute to the dominant row direction of the matrix.

Matrix $\mathbf{M_2}$

$$M_2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 1 & 2 & 0 \\ 0 & 2 & 1 & 2 & 0 \\ 0 & 2 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

- **Rank:** Rows 2, 3, and 4 of M_2 are identical: $[0, 2, 1, 2, 0]$, forming a single linearly independent direction. All other rows are either identical or zero. Thus, the matrix rank is 1.

- **Singular Values:** Singular values correspond to the square roots of the eigenvalues of $M_2 M_2^T$. The product is:

  $$M_2 M_2^T = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 9 & 9 & 9 & 0 \\ 0 & 9 & 9 & 9 & 0 \\ 0 & 9 & 9 & 9 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

  For rank-1 matrices like $M_2 M_2^T$, the only non-zero eigenvalue is equal to the trace of the matrix (as explained in the M_1 case). The trace is:

  $$\text{Trace}(M_2 M_2^T) = 0 + 9 + 9 + 9 + 0 = 27.$$

  Hence, the largest and only non-zero eigenvalue of $M_2 M_2^T$ is 27, and the singular value is:

  $$\sigma_1 = \sqrt{27} = 3\sqrt{3}.$$

- **Left Singular Vector (U):** The left singular vector is derived from the eigenvector of $M_2 M_2^T$. Here, all non-zero rows are proportional to $[0, 1, 1, 1, 0]$. Normalizing this eigenvector gives:

  $$\mathbf{u}_1 = \frac{1}{\sqrt{3}}[0, 1, 1, 1, 0]^\top.$$

- **Right Singular Vector (V):** The right singular vector corresponds to the eigenvector of $M_2^T M_2$. The product $M_2^T M_2$ is:

  $$M_2^T M_2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 12 & 6 & 12 & 0 \\ 0 & 6 & 3 & 6 & 0 \\ 0 & 12 & 6 & 12 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

  The dominant eigenvector is proportional to $[0, 2, 1, 2, 0]$. Normalizing this gives:

  $$\mathbf{v}_1 = \frac{1}{3}[0, 2, 1, 2, 0]^\top.$$

Matrix $\mathbf{M_3}$

$$M_3 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix}$$

- **Rank:** Observing the rows of M_3, rows 2, 3, 4, and 5 are identical: $[0, 1, 1, 1]$, forming a single independent direction. Row 1 is entirely zero, and thus does not contribute to the rank. Therefore, the matrix rank is 1.
- **Singular Values:** As explained in M_1 and M_2, the singular values correspond to the square roots of the eigenvalues of $M_3M_3^T$. The product $M_3M_3^T$ is:

$$M_3M_3^T = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 3 & 3 & 3 \\ 0 & 3 & 3 & 3 & 3 \\ 0 & 3 & 3 & 3 & 3 \\ 0 & 3 & 3 & 3 & 3 \end{bmatrix}.$$

  Similar to M_2, for this rank-1 matrix, the trace gives the only non-zero eigenvalue:

$$\text{Trace}(M_3M_3^T) = 3 + 3 + 3 + 3 + 0 = 12.$$

  The singular value is therefore:

$$\sigma_1 = \sqrt{12} = 2\sqrt{3}.$$

- **Left Singular Vector (U):** The left singular vector is the normalized eigenvector of $M_3M_3^T$. All non-zero rows of M_3 align with the direction $[0, 1, 1, 1, 1]$. Normalizing this vector gives:

$$\mathbf{u}_1 = \frac{1}{2}[0, 1, 1, 1, 1]^\top.$$

- **Right Singular Vector (V):** The right singular vector corresponds to the eigenvector of $M_3^TM_3$. The product $M_3^TM_3$ is:

$$M_3^TM_3 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 4 & 4 & 4 \\ 0 & 4 & 4 & 4 \\ 0 & 4 & 4 & 4 \end{bmatrix}.$$

  The dominant eigenvector of $M_3^TM_3$ is proportional to $[0, 1, 1, 1]$, normalized as:

$$\mathbf{v}_1 = \frac{1}{\sqrt{3}}[0, 1, 1, 1]^\top.$$

Matrix M_4

$$M_4 = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

- **Rank:** Observing M_4, rows 1, 2, and 3 are identical: $[1,1,1,0,0]$, forming one linearly independent direction. Rows 4 and 5 are identical: $[0,0,0,1,1]$, forming another independent direction. Since these two directions are linearly independent of each other, the rank of M_4 is 2.

- **Singular Values:** Singular values correspond to the square roots of the eigenvalues of $M_4 M_4^T$. Since M_4 has rank 2, there are two non-zero singular values. Looking at $M_4 M_4^T$, which is:

  $$M_4 M_4^T = \begin{bmatrix} 3 & 3 & 3 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 2 & 2 \end{bmatrix}.$$

  we can infer that the $\sigma_1 = 3$ and corresponds to the strength (contribution) of the dominant row group, $[1,1,1,0,0]$, and $\sigma_2 = 2$ the second singular value corresponds to the strength of the second row group, $[0,0,0,1,1]$.

- **Left Singular Vectors (U):** The left singular vectors correspond to the normalized eigenvectors of $M_4 M_4^T$. The two non-zero eigenvectors of this product align with the row directions:

  $$\mathbf{u}_1 = \frac{1}{\sqrt{3}}[1,1,1,0,0]^\top, \quad \mathbf{u}_2 = \frac{1}{\sqrt{2}}[0,0,0,1,1]^\top.$$

- **Right Singular Vectors (V):** The right singular vectors correspond to the eigenvectors of $M_4^T M_4$, which is:

  $$M_4^T M_4 = \begin{bmatrix} 3 & 3 & 3 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 2 & 2 \end{bmatrix}.$$

  The two dominant eigenvectors of this product align with the column directions:

  $$\mathbf{v}_1 = \frac{1}{\sqrt{3}}[1,1,1,0,0]^\top, \quad \mathbf{v}_2 = \frac{1}{\sqrt{2}}[0,0,0,1,1]^\top.$$

Matrix $\mathrm{M_5}$

$$M_5 = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

- **Rank:** Observing M_5, rows 1 and 2 are identical: $[1, 1, 1, 0, 0]$, contributing one independent direction. Row 3 introduces a new independent direction: $[1, 1, 1, 1, 1]$, which combines elements from the first two rows and extends to the last two columns. Rows 4 and 5 are identical: $[0, 0, 1, 1, 1]$, forming another independent direction. Since these three row groups are linearly independent, the rank of M_5 is 3.

- **Singular Values:** Singular values correspond to the square roots of the eigenvalues of $M_5 M_5^T$. Since M_5 has rank 3, there are three non-zero singular values. While we cannot compute these explicitly without solving the characteristic equation, we can assume that σ_1 corresponds to the SVD representation of the corresponding row group, $[1, 1, 1, 0, 0]$, σ_2 - to the representation of row 3, $[1, 1, 1, 1, 1]$ and σ_3 - to the representation of rows 4 and 5, $[0, 0, 1, 1, 1]$.

- **Left Singular Vectors (U):** The left singular vectors correspond to the eigenvectors of $M_5 M_5^T$:

  $$M_5 M_5^T = \begin{bmatrix} 3 & 3 & 3 & 1 & 1 \\ 3 & 3 & 3 & 1 & 1 \\ 3 & 3 & 5 & 3 & 3 \\ 1 & 1 & 3 & 3 & 3 \\ 1 & 1 & 3 & 3 & 3 \end{bmatrix}.$$

  This matrix is not orthogonal, which complicates the manual derivation of eigenvectors without additional methods. FAILED

- **Right Singular Vectors (V):** Same as in the point above. FAILED

Matrix $\mathrm{M_6}$

$$M_6 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

- **Rank:** Observing M_6, most rows are identical: $[1, 1, 1, 1, 1]$, forming one independent direction. The third row, $[1, 1, 0, 1, 1]$, introduces a new independent direction since its third column value differs from the other rows. Thus, M_6 has rank 2 because there are two independent directions in the row space.

- **Singular Values:** Singular values correspond to the square roots of the eigenvalues of $M_6M_6^T$. Since M_6 has rank 2, there are two non-zero singular values. While we cannot compute these explicitly without solving the characteristic equation, we infer: - σ_1 corresponds to the dominant row group, $[1, 1, 1, 1, 1]$, - σ_2 corresponds to the independent direction introduced by row 3, $[1, 1, 0, 1, 1]$.

- **Left Singular Vectors (U):** The left singular vectors correspond to the eigenvectors of $M_6M_6^T$:

$$M_6M_6^T = \begin{bmatrix} 5 & 5 & 4 & 5 & 5 \\ 5 & 5 & 4 & 5 & 5 \\ 4 & 4 & 4 & 4 & 4 \\ 5 & 5 & 4 & 5 & 5 \\ 5 & 5 & 4 & 5 & 5 \end{bmatrix}.$$

  Due to the lack of orthogonality in this matrix, it is not possible to derive the eigenvectors manually. Thus, the derivation of the left singular vectors is not feasible without computational tools. FAILED.

- **Right Singular Vectors (V):** The right singular vectors correspond to the eigenvectors of $M_6^TM_6$:

$$M_6^TM_6 = \begin{bmatrix} 5 & 5 & 4 & 5 & 5 \\ 5 & 5 & 4 & 5 & 5 \\ 4 & 4 & 4 & 4 & 4 \\ 5 & 5 & 4 & 5 & 5 \\ 5 & 5 & 4 & 5 & 5 \end{bmatrix}.$$

  Similar to $M_6M_6^T$, the lack of orthogonality makes it impossible to manually compute the eigenvectors of this matrix. Thus, the derivation of the right singular vectors is also not feasible without computational tools. FAILED.

B. Visualizations of Best Rank-1 Approximations

C. Feature Distribution

Figure **??** shows the distribution of the features after normalization. Although the scale has been changed by normalization, the overall shaped of the distribution remains the same, centered around mean at 0.

D. Fragments SVD components

Figure 10 shows the first five rows of the matrix Vt - the first five right singular vectors of the data.

Figure 11 shows the first ten rows and columns of the matrix S, whose diagonal represents the first ten singular values of the data.

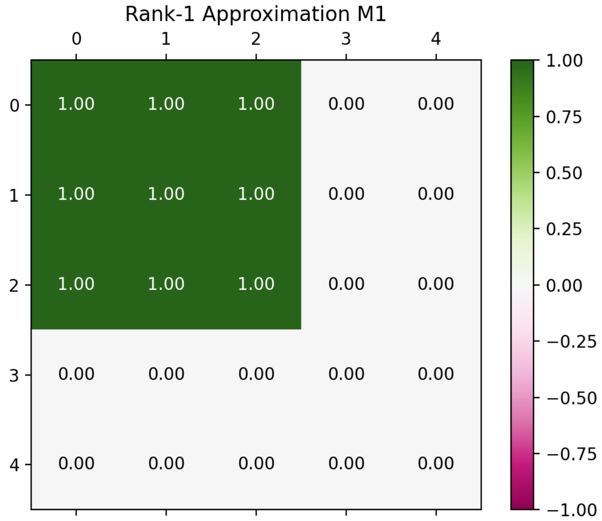


Figure 3: Best rank-1 approximation for M_1.

E. Interpretation of the Columns of U Based on North-South and East-West Distinction

Figure 12 shows the data points colored according to their North-South location with the first left singular vector values on the x-axis and the second to fifth left singular vectors values on the y-axis on each plot respectively.

Figure 13 shows the data points colored according to their East-West location with the first left singular vector values on the x-axis and the second to fifth left singular vectors values on the y-axis on each plot respectively.

Figure 14 shows the data points colored according to their North-South and East-West location with the second left singular vector values on the x-axis and the third left singular vectors values on the y-axis.

F. Selection of the rank value for Truncated SVD

Figure 15 shows the Cattell's Scree test performed to define the optimal k for truncated SVD. The plot also contains the result of a KneeLocator algorithm applied to verify our assumption.

Figure 16 shows the selection of k for truncated SVD with random flip of sign coin, where we select the value of k that provides the smallest change in the spectral norm of the original and randomly perturbed residual matrices.

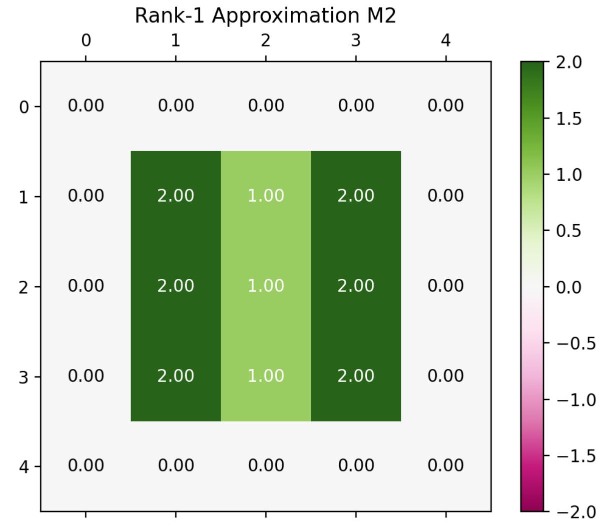


Figure 4: Best rank-1 approximation for M_2.

G. K-Means Clustering of the Climate Dataset

Figure 17 shows the geographical location of the data points colored according to the cluster assigned to them by K-Means clustering with K=5.

Figure 18 shows the data points colored according to the cluster assigned to them by K-Means clustering with K=5 represented by their values of the first left singular vector on the x-axis and the second left singular value on the y-axis.

Figure 19 shows the data points colored according to the cluster assigned to them by K-Means clustering with K=7 represented by their values of the first left singular vector on the x-axis and the second left singular value on the y-axis.

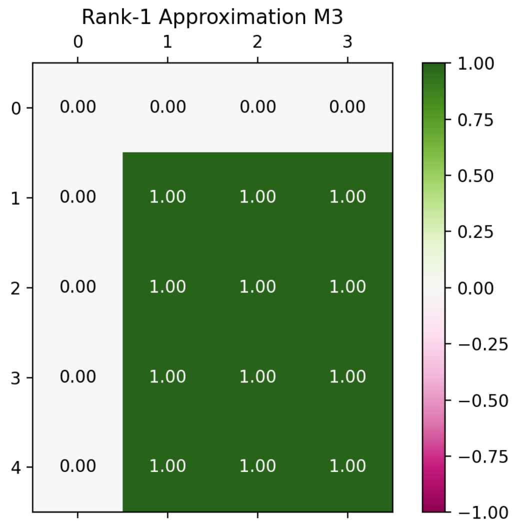


Figure 5: Best rank-1 approximation for M_3.

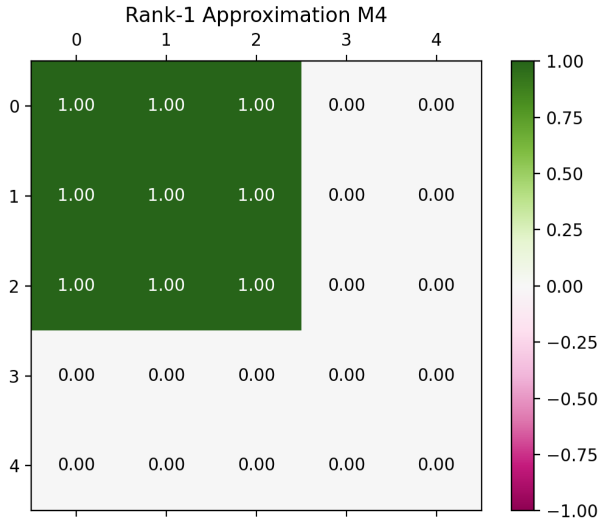


Figure 6: Best rank-1 approximation for M_4.

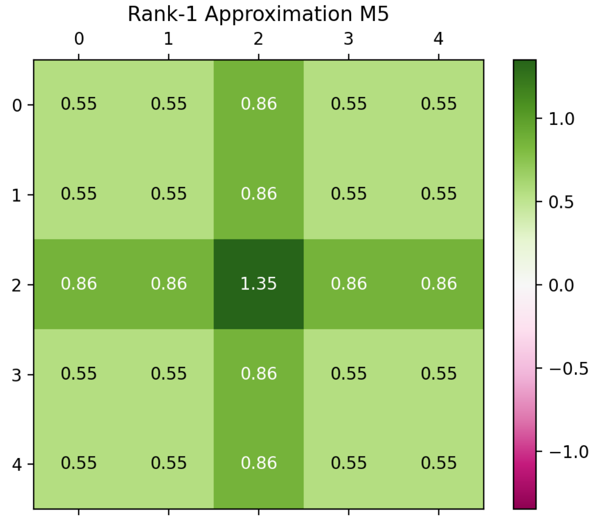


Figure 7: Best rank-1 approximation for M_5.

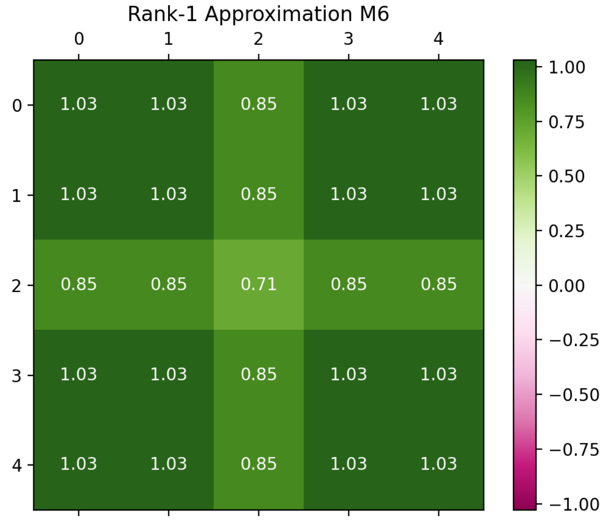


Figure 8: Best rank-1 approximation for M_6.

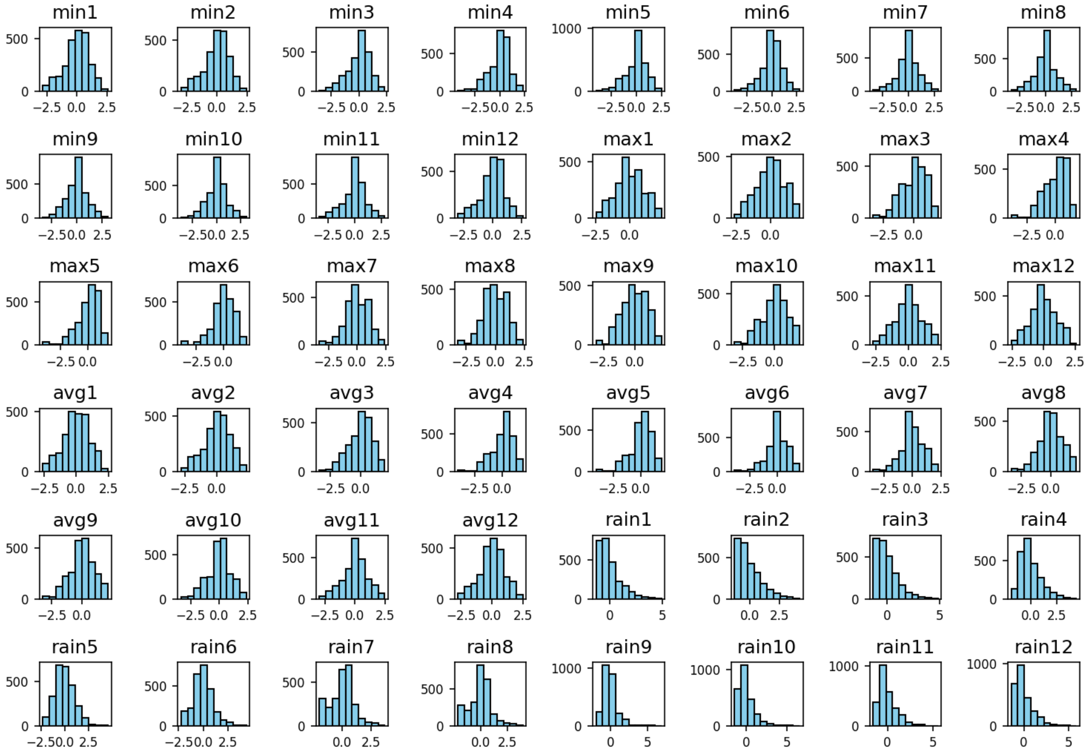


Figure 9: Distribution of features

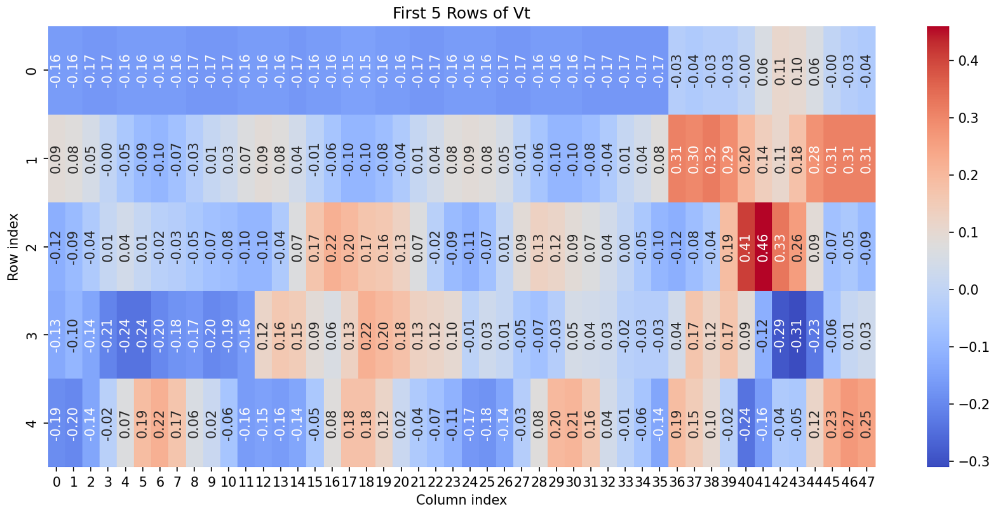


Figure 10: The first five rows of the matrix Vt (rounded to two decimal points)

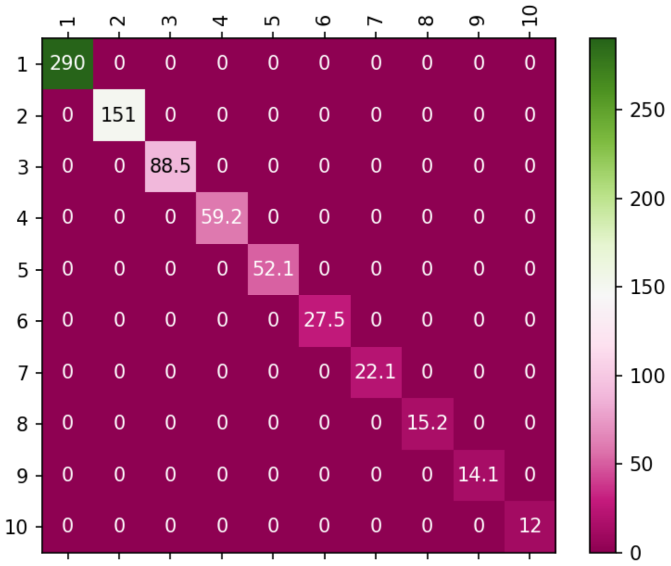


Figure 11: The first ten singular values)

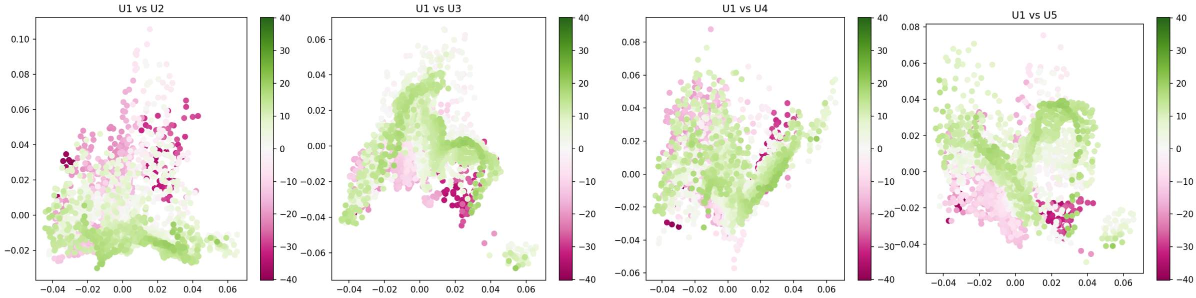


Figure 12: The First Column of U against the following four columns for North-South location)

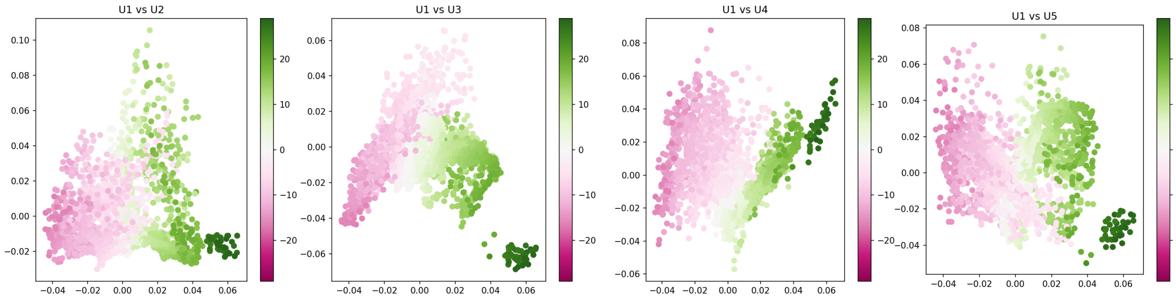


Figure 13: The First Column of U against the following four columns for East-West location)

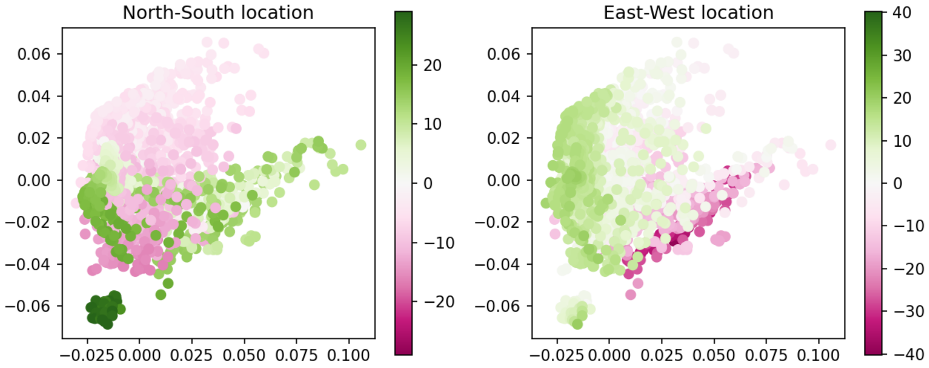


Figure 14: The Second Column of U against the Third Column of U)

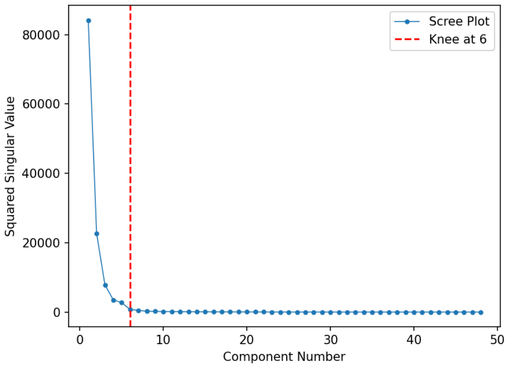


Figure 15: Cattell's Scree test with the KneeLocator algorithm

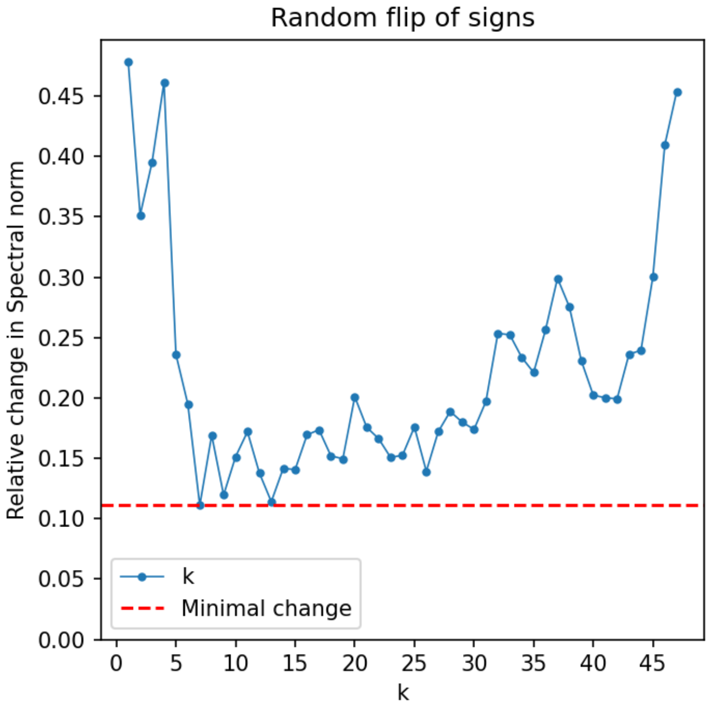


Figure 16: Random flip of signs method for selection of k

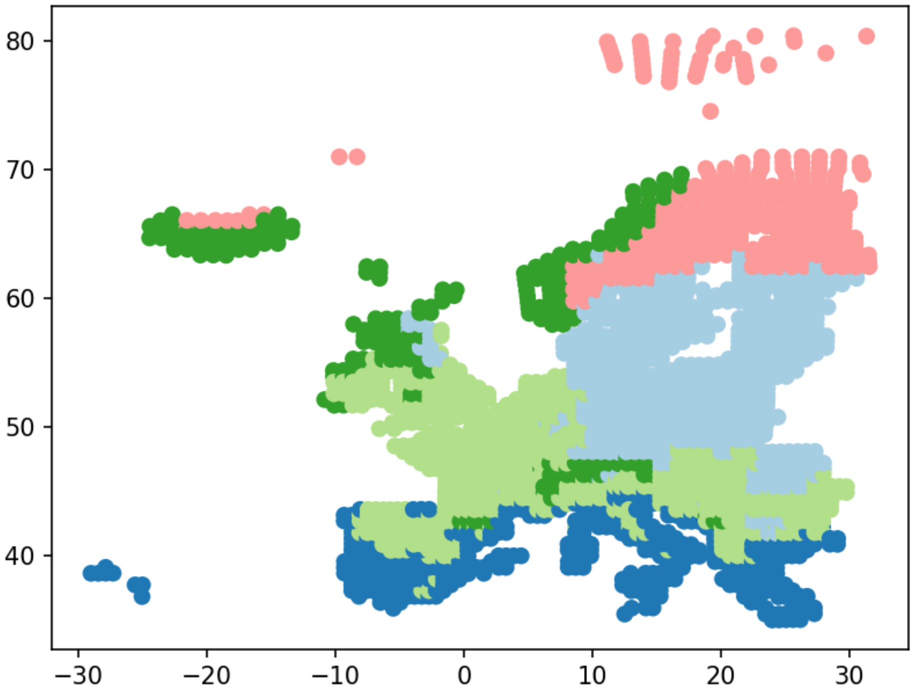


Figure 17: K-Means Clustering of the Climate Data

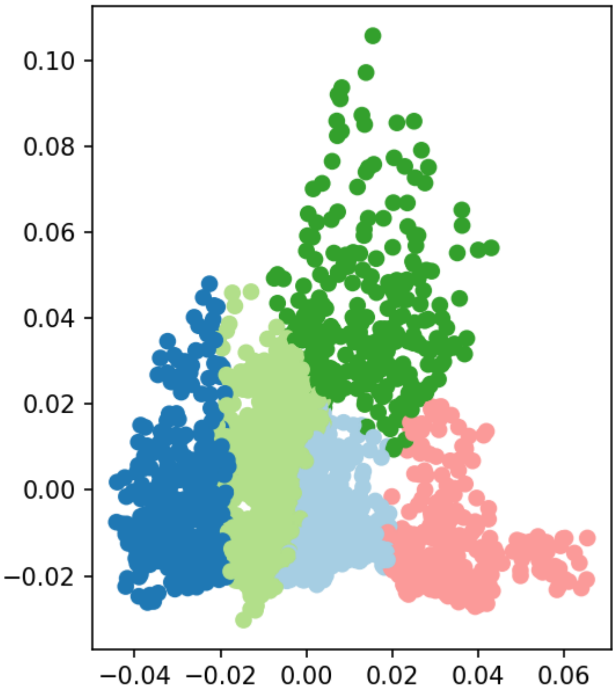


Figure 18: K-Means Clusters in the Singular Vector Space (K=5)

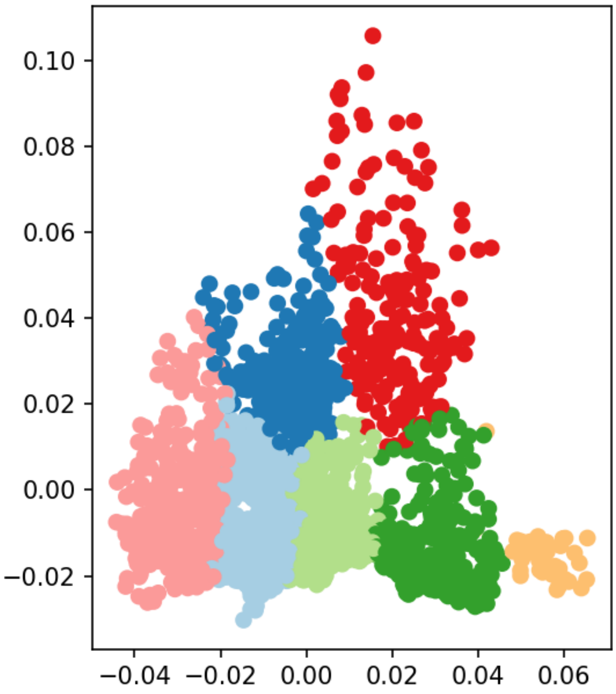


Figure 19: K-Means Clusters in the Singular Vector Space (K=7)

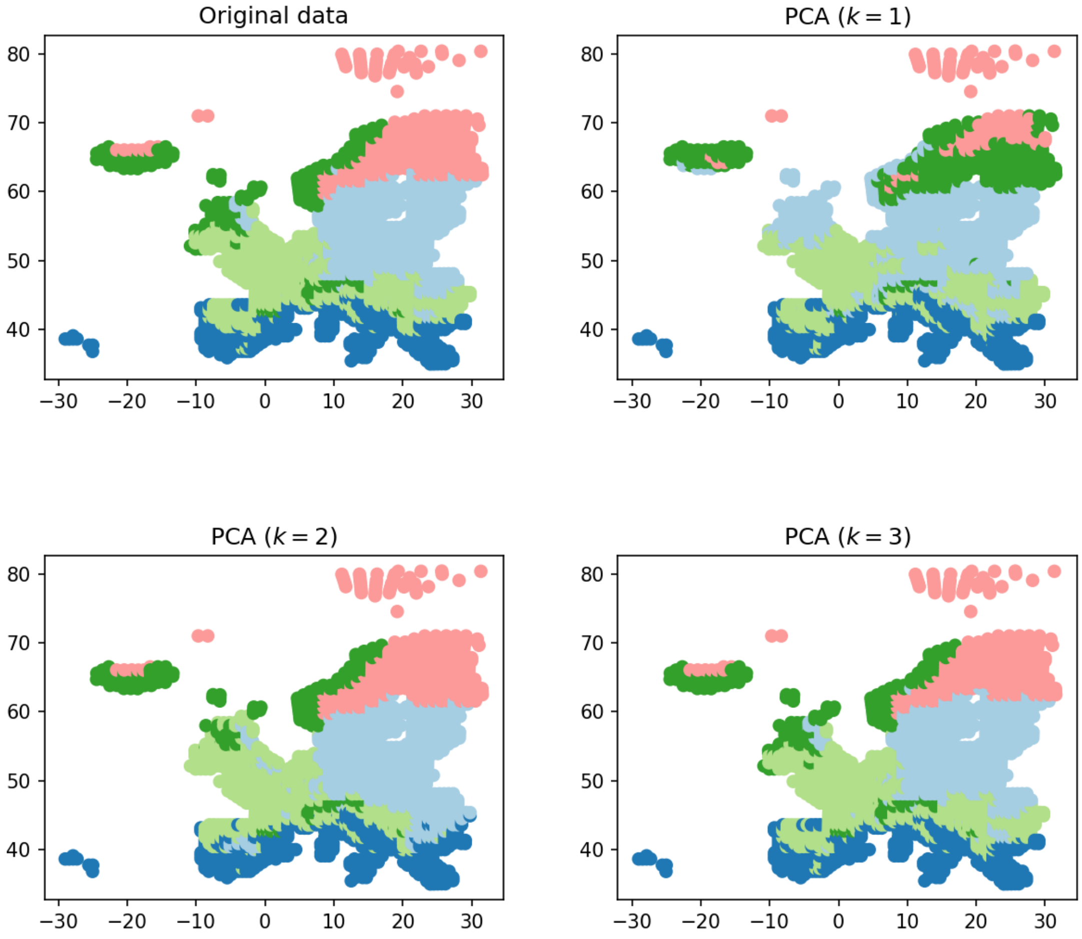


Figure 20: Clustering of original data vs. PCA

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	Style Edits	Throughout	++
GPT-4	Code debugging	Throughout	+-

Unterschrift
Mannheim, den 17. November 2024