Laboratory work 9: Dimension reduction with SVD

Olha Makovlieva, IKM-M225d

Tasks

For the corresponding dataset, reduce the dimensionality of the data using PCA and SVD.

1. Using **PCA** to visualize data in two- and three-dimensional (**2D and 3D**) spaces.
2. Calculate **SVD** of your dataset, plot the dependence of the eigenvalues of the matrix on their number. Before plotting, arrange the eigenvalues in descending order.
3. Determine the smallest value of the space size d for which the following relation is satisfied:

$$\frac{\sum_{i=0}^{d} \lambda_i}{\sum_{i=0}^{n} \lambda_i} \geq 0.8,$$

   where λ_i are the eigenvalues of the matrix, n is the total number of eigenvalues, and 0.8 is the level of data significance.
4. Set λ_i to zero for $d \leq i \leq n$. Perform the reverse transformation and compare the obtained data with the original.
5. Set $d = 2$ (for 2D) and $d = 3$ (for 3D), perform and plot the first d columns of reconstructed data. Compare the graph with the one obtained in step 1.

Results

PCA Visualization

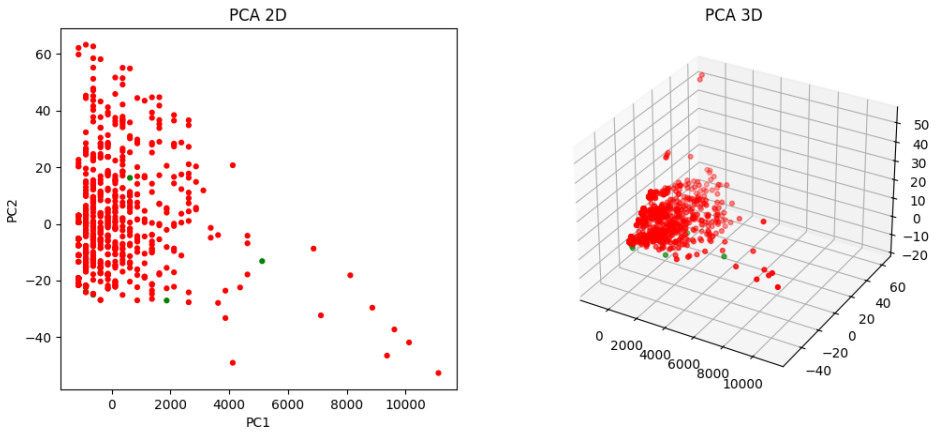


Figure 1: PCA 2D and 3D Visualization

In the **2D PCA plot**, the data points are displayed based on the two directions of maximum variance, allowing visual inspection of whether the classes form distinguishable patterns or clusters. The **3D PCA**

plot extends this by adding a third component, capturing more variance and potentially revealing structure not visible in two dimensions. The color coding reflects the class labels, allowing comparison of how well-separated the groups appear in the reduced-dimensional space.

Eigenvalue Analysis

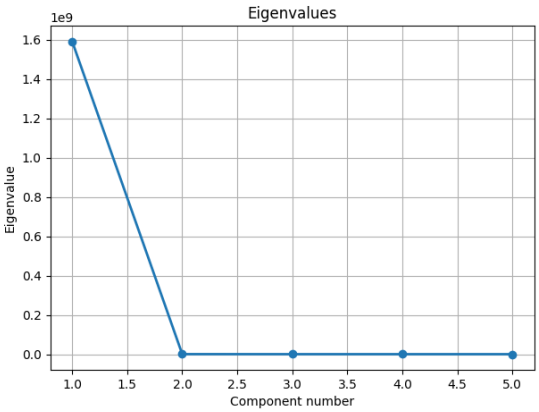


Figure 2: Eigenvalues

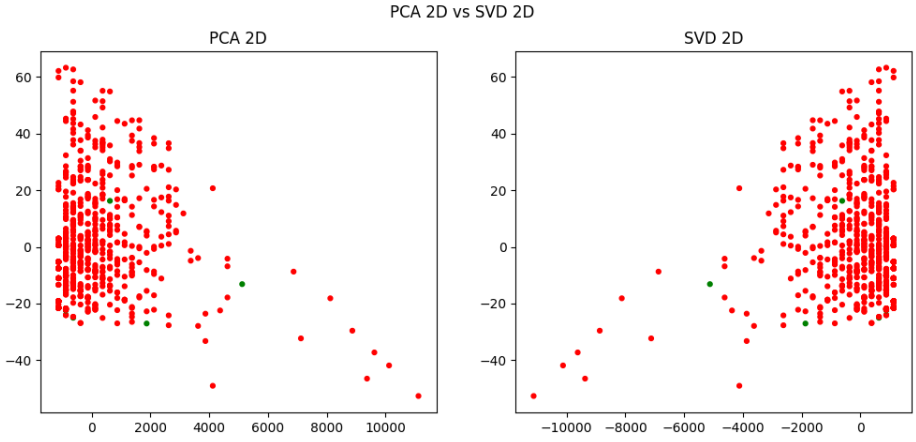
The eigenvalue graph displays the sorted squared singular values obtained from the Singular Value Decomposition of the centered dataset. Each eigenvalue represents the amount of variance captured by its corresponding principal component. Larger eigenvalues indicate components that explain more variability in the data. The downward-sloping curve helps identify how quickly the variance decreases and whether there is a natural cutoff point where additional components contribute very little to the total information.

PCA vs. SVD Comparison

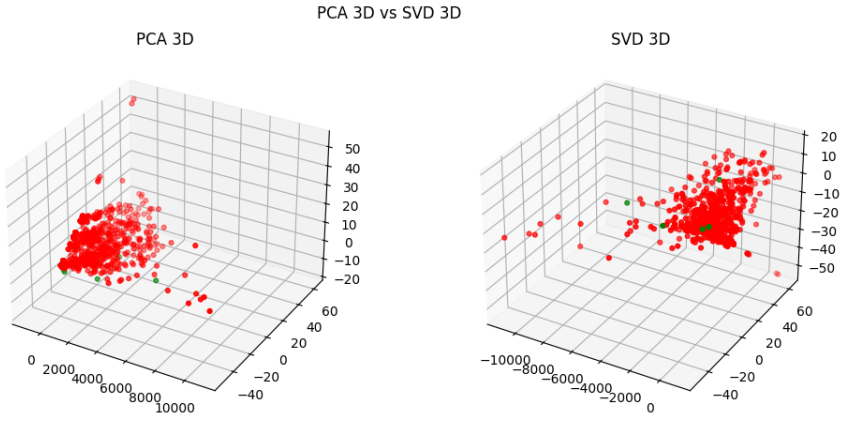
The comparison plots demonstrate that **PCA and SVD yield equivalent low-dimensional representations** when applied to centered data. The 2D and 3D scatterplots for PCA and SVD show nearly identical spatial distributions of points, confirming the mathematical equivalence of the two methods for dimensionality reduction. Any visual differences between the plots are due only to rotation or scaling conventions, but the underlying relationships between the data points remain the same.

Optimal Dimensionality

Optimal d by mean squared error (MSE) is 3, with MSE = 0.0316. For each possible dimension d, the algorithm reconstructs the data using only the first d singular values and computes the **mean squared reconstruction error**. The smallest value of d that keeps the error below the chosen threshold is reported as the optimal dimensionality. This result indicates how many components are sufficient to preserve the essential structure of the dataset.



(a) PCA 2D vs SVD 2D



(b) PCA 3D vs SVD 3D

Figure 3: Comparison of PCA and SVD Low-Dimensional Representations