In-class Exercise 15.1
Singular value decomposition
AMATH 301
University of Washington
Jakob Kotas

1. Consider the square matrix:

$$A = \begin{bmatrix} -11 & 45 & -33 \\ 7 & 2 & -25 \\ 0 & -8 & 4 \end{bmatrix}.$$

   (a) Use `np.linalg.svd` to find the Singular Value Decomposition of A.

   (b) Verify that $A = U\Sigma V^T$.

   (c) Verify that the eigenvalues of AA^T and A^TA are the same, and that their square roots are the singular values contained in Σ.

   (d) Verify that the columns of U are the eigenvectors of AA^T and the columns of V are the eigenvectors of A^TA.

2. Repeat #1 (a)-(d) with the non-square matrix:

$$A = \begin{bmatrix} -11 & 45 & -33 \\ 7 & 2 & -25 \\ 0 & -8 & 4 \\ 1 & 1 & 1 \end{bmatrix}.$$

3. SVD has very important applications in statistics and machine learning, where it is used for principal component analysis (PCA) and dimension reduction.

   Below are genetic data on six patients. Y_1, Y_2, and Y_3 have a certain disease and N_1, N_2, and N_3 do not. The expression of 8 genes were measured in each of the 6 patients.

Patient	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7	Gene 8
Y_1	0.98	1.00	0.99	0.02	0.01	0.03	0.05	0.98
Y_2	0.99	0.97	0.96	0.01	0.04	0.02	0.01	0.99
Y_3	0.96	0.99	0.92	0.05	0.02	0.02	0.02	0.96
N_1	0.02	0.01	0.03	0.91	0.98	0.06	0.01	0.96
N_2	0.03	0.08	0.04	0.99	0.94	0.01	0.04	0.99
N_3	0.00	0.02	0.05	0.96	0.98	0.02	0.04	0.99

   (a) From inspection, can you tell if there are certain genes that tend to be higher or lower in patients with the disease compared to patients without?

   (b) Find the SVD using Python and show that most of the information (variance) is given by the first two singular values.

   (c) Next we will perform dimensionality reduction, so we will only consider the two largest singular values (call this diagonal matrix S_2) and the corresponding first two columns in the U matrix (call this U_2). Create the matrix $B = U_2S_2$. B is size 6×2.

   (d) Plot the coordinates in B on xy-axes. This is a 2D projection of the 6D data we began with. Does the figure make sense? If a 7th patient came along with data on these 8 genes, could we predict whether they have the disease?

   (e) Let the first two rows in the V^T matrix be V_2^T. Create the matrix $C = S_2V_2$. C is size 2×8. Also plot the coordinates in C on xy-axes. Can we see which genes tend to be correlated with each other?

4. Below is some climatic data from 18 US cities.

City	Precip (in.)	Precip (days)	July High (F)	Jan High (F)	Annual Sunshine (hr)
Seattle	39.34	156.2	77.4	48	2169.7
Portland	36.92	157	81.9	47.5	2340.9
Juneau	66.99	230.2	64	33.1	1530.7
San Francisco	22.89	71.2	66.3	57.8	3061.7
Los Angeles	14.25	34.1	82	68	3254.2
Phoenix	7.22	33.4	106.5	67.6	3871.6
Las Vegas	4.18	25.8	104.5	58.5	3825.3
Miami	67.41	141	90.6	76.2	3154
Honolulu	16.41	89.2	88.1	80.5	3035.9
Hilo	120.39	273	82.8	78.7	1817.4
Chicago	40.88	127	85.2	32.8	2508.4
New York City	49.52	125.4	84.9	39.5	2534.7
Anchorage	16.42	115.1	66.2	22.7	2061.2
Fairbanks	11.67	107.1	72.7	-0.6	2105
New Orleans	63.35	115.1	91.4	62.5	2648.9
Minneapolis	31.62	118.8	83.4	23.6	2710.7
Denver	14.48	79.7	89.9	44.6	3106.6
Boise	11.51	89.2	92.7	38.8	2993.4

(a) Create a matrix A containing these values.

(b) Scale each column of A so that its mean is 0 and standard deviation is 1. In statistics, this is called the z-score of each value: we care about how many standard deviations above or below the mean each data point is.

(c) Perform SVD on the matrix.

(d) Next we will perform dimensionality reduction, so we will only consider the two largest singular values (call this diagonal matrix S_2) and the corresponding first two columns in the U matrix (call this U_2). Create the matrix $B = U_2 S_2$. B is size 18×2.

(e) Plot the coordinates in B on xy-axes. This is a 2D projection of the 5D data we began with. Can we make any sense of the figure?