DSC5001: Statistical Machine Learning I

Tutorial 12: PCA, SVD, and K-Means Clustering

Kangning Cui
City University of Hong Kong (Dongguan)

Semester A, 2026/27

Task 1: Sensor Compression

Four observations have two sensor readings:

Case	A	B	C	D
x_1	4	-4	1	-1
x_2	2	-2	-2	2

Use covariance $\boldsymbol{S} = \frac{1}{n}\widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}}$, where $n = 4$.

(a) Find the sample mean and the 2×2 covariance matrix $\boldsymbol{S}$.

(b) Find both eigenvalues and corresponding unit eigenvectors. Identify PC1.

(c) Compute the PC1 scores of all four cases and the proportion of variance explained by PC1. Which cases become indistinguishable in one dimension?

(d) For a new, already-centered observation $\boldsymbol{x}_* = (3, 1)^\top$, calculate its PC1 score, one-PC reconstruction, and squared reconstruction error.

Solution 1(a): Centering

Compute the mean of each feature:

$$\bar{x}_1 = \frac{4 - 4 + 1 - 1}{4} = 0, \qquad \bar{x}_2 = \frac{2 - 2 - 2 + 2}{4} = 0.$$

The data are already centered:

$$\boxed{\bar{\boldsymbol{x}} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \widetilde{\boldsymbol{X}} = \begin{pmatrix} 4 & 2 \\ -4 & -2 \\ 1 & -2 \\ -1 & 2 \end{pmatrix}.}$$

Solution 1(a): Covariance Matrix

Using $\boldsymbol{S} = \widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}}/4$,

$$S_{11} = \frac{16+16+1+1}{4} = \frac{17}{2},$$
$$S_{22} = \frac{4+4+4+4}{4} = 4,$$
$$S_{12} = \frac{8+8-2-2}{4} = 3.$$

Therefore,

$$\boxed{\boldsymbol{S} = \begin{pmatrix} 17/2 & 3 \\ 3 & 4 \end{pmatrix}.}$$

Solution 1(b): Eigenvalues

Solve $\det(\boldsymbol{S} - \lambda \boldsymbol{I}) = 0$:

$$\begin{aligned} 0 &= \left(\frac{17}{2} - \lambda\right)(4 - \lambda) - 9 \\ &= \lambda^2 - \frac{25}{2}\lambda + 25 \\ &= (\lambda - 10)\left(\lambda - \frac{5}{2}\right). \end{aligned}$$

Thus

$$\boxed{\lambda_1 = 10, \qquad \lambda_2 = \frac{5}{2}.}$$

Solution 1(b): Principal Directions

For $\lambda_1 = 10$, solving $(\boldsymbol{S} - 10\boldsymbol{I})\boldsymbol{v} = 0$ gives $v_1 = 2v_2$.
For $\lambda_2 = 5/2$, the eigenvector has $v_2 = -2v_1$.
Normalize both directions:

$$\boxed{\boldsymbol{v}_1 = \frac{1}{\sqrt{5}}\begin{pmatrix}2\\1\end{pmatrix}, \qquad \boldsymbol{v}_2 = \frac{1}{\sqrt{5}}\begin{pmatrix}1\\-2\end{pmatrix}.}$$

Check:

$$\boldsymbol{v}_1^\top \boldsymbol{v}_2 = \frac{2-2}{5} = 0.$$

Solution 1(c): Principal Scores

Project each observation onto the first eigenvector:

$$z_{i1} = \boldsymbol{v}_1^\top \boldsymbol{x}_i = \frac{2x_{i1} + x_{i2}}{\sqrt{5}}.$$

Case	A	B	C	D
$2x_1 + x_2$	10	-10	0	0
PC1 score	$2\sqrt{5}$	$-2\sqrt{5}$	0	0

$$\boxed{z_{C,1} = z_{D,1} = 0.}$$

Solution 1(c): Explained Variance

Total variance:

$$\lambda_1 + \lambda_2 = 10 + \frac{5}{2} = \frac{25}{2}.$$

Variance explained by PC1:

$$\boxed{\mathrm{EVR}(1) = \frac{10}{25/2} = \frac{4}{5} = 80\%.}$$

Discarded variance:

$$\boxed{1 - \mathrm{EVR}(1) = 20\%.}$$

Cases C and D both have PC1 score 0.

Solution 1(d): New-Case Score

For

$$\boldsymbol{x}_* = (3,1)^\top, \qquad \boldsymbol{v}_1 = (2,1)^\top/\sqrt{5},$$

calculate

$$\begin{aligned} z_{*,1} &= \boldsymbol{v}_1^\top \boldsymbol{x}_* \\ &= \frac{2(3)+1(1)}{\sqrt{5}} \\ &= \boxed{\frac{7}{\sqrt{5}}}. \end{aligned}$$

Solution 1(d): Reconstruction

Reconstruct using only PC1:

$$\begin{aligned}\widehat{\boldsymbol{x}}_* &= z_{*,1}\boldsymbol{v}_1 \\ &= \frac{7}{\sqrt{5}}\frac{1}{\sqrt{5}}\begin{pmatrix}2\\1\end{pmatrix} \\ &= \boxed{\begin{pmatrix}14/5\\7/5\end{pmatrix}}.\end{aligned}$$

Solution 1(d): Reconstruction Error

Residual:

$$\begin{aligned} \boldsymbol{e}_* &= \boldsymbol{x}_* - \widehat{\boldsymbol{x}}_* \\ &= \begin{pmatrix} 3 \\ 1 \end{pmatrix} - \begin{pmatrix} 14/5 \\ 7/5 \end{pmatrix} \\ &= \boxed{\begin{pmatrix} 1/5 \\ -2/5 \end{pmatrix}}. \end{aligned}$$

Squared reconstruction error:

$$\boxed{\|\boldsymbol{e}_*\|_2^2 = \frac{1}{25} + \frac{4}{25} = \frac{1}{5} = 0.2.}$$

Task 2: SVD and Data Compression

A centered training matrix has $n = 6$ observations and $p = 3$ features.
Its singular values are $d_1 = 10, d_2 = 6, d_3 = 2$. The first two right singular vectors are

$$\boldsymbol{v}_1 = \frac{1}{\sqrt{2}}(1, 1, 0)^\top, \qquad \boldsymbol{v}_2 = (0, 0, 1)^\top.$$

Use covariance $\boldsymbol{S} = \widetilde{\boldsymbol{X}}^\top \widetilde{\boldsymbol{X}}/6$.

(a) Calculate the covariance eigenvalues and total variance.

(b) Find the smallest number of PCs retaining at least 90% of training variance.

(c) For the rank-two truncated SVD, calculate the total squared training reconstruction error and the average squared error per training observation.

(d) A new centered observation is $\boldsymbol{x}_* = (3, -1, 2)^\top$. Find its first two scores, rank-two reconstruction, and squared error. Compare with part (c).

Solution 2(a): Covariance Eigenvalues

Use $\lambda_j = d_j^2/n$ with $n = 6$:

$$\lambda_1 = \frac{10^2}{6} = \boxed{\frac{50}{3}},$$

$$\lambda_2 = \frac{6^2}{6} = \boxed{6},$$

$$\lambda_3 = \frac{2^2}{6} = \boxed{\frac{2}{3}}.$$

Total variance:

$$\boxed{\sum_{j=1}^{3} \lambda_j = \frac{100 + 36 + 4}{6} = \frac{70}{3}.}$$

Solution 2(b): Component Count

The explained-variance ratio uses squared singular values:

$$\mathrm{EVR}(r) = \frac{\sum_{j=1}^{r} d_j^2}{100 + 36 + 4}.$$

Retained PCs	1	2	3
Retained energy	100	136	140
Cumulative EVR	$5/7$	$34/35$	1

Since

$$\frac{5}{7} \approx 71.4\%, \qquad \frac{34}{35} \approx 97.1\%,$$

$\boxed{r = 2 \text{ is the smallest choice reaching } 90\%.}$

Solution 2(c): Truncated-SVD Error

Rank two keeps d_1, d_2 and discards $d_3 = 2$.
Total squared reconstruction error:

$$\boxed{\|\widetilde{\boldsymbol{X}} - \widetilde{\boldsymbol{X}}_2\|_F^2 = d_3^2 = 4.}$$

Average squared error per training observation:

$$\boxed{\frac{1}{6}\sum_{i=1}^{6} \|\widetilde{\boldsymbol{x}}_i - \widehat{\boldsymbol{x}}_i\|_2^2 = \frac{4}{6} = \frac{2}{3}.}$$

Solution 2(d): New-Case Scores

For $\boldsymbol{x}_* = (3, -1, 2)^\top$,

$$
\begin{aligned}
z_{*,1} &= \boldsymbol{v}_1^\top \boldsymbol{x}_* \\
&= \frac{3-1}{\sqrt{2}} \\
&= \boxed{\sqrt{2}},
\end{aligned}
$$

$$
\begin{aligned}
z_{*,2} &= \boldsymbol{v}_2^\top \boldsymbol{x}_* \\
&= \boxed{2}.
\end{aligned}
$$

Solution 2(d): New-Case Reconstruction

Retain the first two directions:

$$\begin{aligned}\widehat{\boldsymbol{x}}_* &= z_{*,1}\boldsymbol{v}_1 + z_{*,2}\boldsymbol{v}_2 \\ &= \begin{pmatrix}1\\1\\0\end{pmatrix} + \begin{pmatrix}0\\0\\2\end{pmatrix} \\ &= \boxed{\begin{pmatrix}1\\1\\2\end{pmatrix}}.\end{aligned}$$

Residual and squared error:

$$\boxed{\boldsymbol{e}_* = (2,-2,0)^\top, \qquad \|\boldsymbol{e}_*\|_2^2 = 8.}$$

Solution 2(d): Error Comparison

From part (c), the average squared training error was

$$\frac{2}{3}.$$

For the new observation,

$$\boxed{\|\boldsymbol{e}_*\|_2^2 = 8 > \frac{2}{3}.}$$

The two retained PCs account for

$$\boxed{\mathrm{EVR}(2) = \frac{34}{35} \approx 97.1\%}$$

of the training variance.

Task 3: Machine Vibration Clustering

Six machines have one vibration reading (mm/s):

Machine	A	B	C	D	E	F
x	1	3	4	7	11	12

Run $K = 2$ K-means, starting at $\mu_1^{(0)} = 0$ and $\mu_2^{(0)} = 6$.
Assign exact ties to cluster 1. Use $J = \sum_i (x_i - \mu_{c_i})^2$.

(a) Find the initial assignments and J_0.

(b) Update both centroids without changing assignments. Find the new objective.

(c) Reassign all points, update the centroids again, and compute the final objective. Verify that assignments are stable.

(d) For the observation $x = 7$ in the final partition, calculate its silhouette using *ordinary absolute distance*. Does K-means convergence guarantee a positive silhouette?

Solution 3(a): Initial Assignments

Initial centers:

$$\mu_1^{(0)} = 0, \qquad \mu_2^{(0)} = 6.$$

Their midpoint is

$$\frac{0+6}{2} = 3.$$

Assign the tie at $x = 3$ to cluster 1:

$$\boxed{C_1^{(0)} = \{1, 3\}, \qquad C_2^{(0)} = \{4, 7, 11, 12\}.}$$

Solution 3(a): Initial Objective

Using the initial centers 0 and 6,

$$\begin{aligned} J_0 &= (1-0)^2 + (3-0)^2 \\ &\quad + (4-6)^2 + (7-6)^2 \\ &\quad + (11-6)^2 + (12-6)^2 \\ &= 1 + 9 + 4 + 1 + 25 + 36 \\ &= \boxed{76}. \end{aligned}$$

Solution 3(b): First Centroid Update

Update the means, keeping assignments fixed:

$$\boxed{\mu_1^{(1)} = \frac{1+3}{2} = 2, \qquad \mu_2^{(1)} = \frac{4+7+11+12}{4} = \frac{17}{2}.}$$

The within-cluster squared errors are 2 and 41:

$$\begin{aligned} J &= \left[(-1)^2 + 1^2\right] \\ &\quad + \left[(-4.5)^2 + (-1.5)^2 + (2.5)^2 + (3.5)^2\right] \\ &= 2 + 41 \\ &= \boxed{43}. \end{aligned}$$

Solution 3(c): Reassignment

The current centers are 2 and 8.5; their midpoint is 5.25.

$$\boxed{C_1^{(1)} = \{1, 3, 4\}, \qquad C_2^{(1)} = \{7, 11, 12\}.}$$

With these assignments and the same centers,

$$\begin{aligned} J &= [1^2 + 1^2 + 2^2] \\ &\quad + [(-1.5)^2 + (2.5)^2 + (3.5)^2] \\ &= 6 + \frac{83}{4} \\ &= \boxed{\frac{107}{4} = 26.75}. \end{aligned}$$

Solution 3(c): Second Update

New centroids:

$$\boxed{\mu_1^{(2)} = \frac{1+3+4}{3} = \frac{8}{3}, \qquad \mu_2^{(2)} = \frac{7+11+12}{3} = 10.}$$

Within-cluster errors:

$$J_{C_1} = \frac{25+1+16}{9} = \frac{14}{3},$$
$$J_{C_2} = (-3)^2 + 1^2 + 2^2 = 14.$$

Total:

$$\boxed{J_{\text{final}} = \frac{14}{3} + 14 = \frac{56}{3} \approx 18.67.}$$

Solution 3(c): Convergence Check

The new centers are $8/3$ and 10. Their midpoint is

$$\boxed{\frac{1}{2}\left(\frac{8}{3}+10\right)=\frac{19}{3}\approx 6.33.}$$

Every reading in cluster 1 satisfies

$$x \le 4 < 19/3.$$

Every reading in cluster 2 satisfies

$$x \ge 7 > 19/3.$$

Thus the final assignments are unchanged:

$$\boxed{C_1=\{1,3,4\}, \qquad C_2=\{7,11,12\}.}$$

Solution 3(d): Silhouette Distances

For $x = 7$, its own cluster is $\{7, 11, 12\}$. Use ordinary absolute distances.
Average within-cluster distance:

$$\boxed{a(7) = \frac{|7-11| + |7-12|}{2} = \frac{9}{2}.}$$

Average distance to the other cluster $\{1, 3, 4\}$:

$$\boxed{b(7) = \frac{|7-1| + |7-3| + |7-4|}{3} = \frac{13}{3}.}$$

Solution 3(d): Silhouette Score

Apply the silhouette formula:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}.$$

For $x = 7$,

$$\begin{aligned} s(7) &= \frac{13/3 - 9/2}{9/2} \\ &= \frac{-1/6}{9/2} \\ &= \boxed{-\frac{1}{27} \approx -0.037}. \end{aligned}$$

Answer: No. K-means has converged, but $s(7) < 0$.

Tutorial 12: PCA, SVD, and K-means

Questions and discussion