STA 35C – Homework 7

Submission due: Thu, December 3 at 11:59 PM PT

Instructor: Dogyoon Song

Instructions: Upload a **single PDF file** to Gradescope via Canvas ("Homework 7" under "Assignments"). Name it using your UC Davis email ID prefix and homework number, e.g., `dgsong_hw7.pdf`. Include "STA 35C," your name, and the last four digits of your student ID on the front page. For coding problems, include your R code, relevant output, and requested plots/tables in the PDF. Late submissions receive 0 points. See the syllabus for full submission and late-policy details. Complete Problems 1–3 (**100 points total**). Problem 4 offers two optional parts worth **5 bonus points each**. You may use R throughout. Show the intermediate calculations specifically requested.

Problem 1: Principal component analysis (40 points total).

You have a *two-dimensional* dataset of 10 points, $\{x_1, \ldots, x_{10}\} \subset \mathbb{R}^2$. Suppose the 10 data points are:

$$x_1 = (2,\ 3), \quad x_2 = (2,\ 5), \quad x_3 = (3,\ 6), \quad x_4 = (4,\ 5), \quad x_5 = (5,\ 8),$$
$$x_6 = (6,\ 10), \quad x_7 = (6,\ 7), \quad x_8 = (7,\ 11), \quad x_9 = (7,\ 9), \quad x_{10} = (8,\ 12).$$

(a) (10 points)

(i) Compute the sample mean $\bar{x} \in \mathbb{R}^2$ and *center* the data by subtracting $\bar{x}$ from each point.

(ii) Compute the 2×2 *empirical covariance matrix*:

$$\Sigma \ = \ \frac{1}{10} \sum_{i=1}^{10} (x_i - \bar{x}) \, (x_i - \bar{x})^\top.$$

(*Note:* Typically, we treat each data point x_i as a (column) vector, e.g., $x_1 = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$. Then its transpose $x_i^\top$ is the row vector, e.g., $x_1^\top = \begin{bmatrix} 2 & 3 \end{bmatrix}$. You can compute the product $\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \begin{bmatrix} b_1 & b_2 \end{bmatrix} = \begin{bmatrix} a_1 b_1 & a_1 b_2 \\ a_2 b_1 & a_2 b_2 \end{bmatrix}$. In **R**, you can compute such an outer product by `a %*% t(b)`, where `a` and `b` are column vectors.)

(iii) Using R (e.g. `prcomp` or an eigen-decomposition of Σ), find the *first principal component direction* $\mathbf{u}_1$. (*Note:* The sign of $\mathbf{u}_1$ as a principal component is arbitrary, so $\mathbf{u}_1$ and $-\mathbf{u}_1$ are equivalent.)

(b) (10 points)

(i) Compute the *directional variance* along $\mathbf{e}_1 = (1, 0)$, i.e. the variance of $\langle \mathbf{e}_1, x_i \rangle$.

(ii) Compute the directional variance along your first principal component $\mathbf{u}_1$ from part **(a)**(iii).

(iii) Verify that these directional variances match $\mathbf{e}_1^\top \, \Sigma \, \mathbf{e}_1$ and $\mathbf{u}_1^\top \, \Sigma \, \mathbf{u}_1$ for Σ computed above, respectively.

(c) (10 points) In R, generate a *synthetic* dataset with $p = 25$ variables and $n = 100$ observations. For example, use the following code snippet (mixture of 4 Gaussians, with two means close together):

```
set.seed(2026)
n  <- 100
p  <- 25

mean1 <- c(2, rep(0, p-1))
mean2 <- c(0, 2, rep(0, p-2))
mean3 <- c(1, 1.5, -1, rep(0, p-3))
mean4 <- c(1, 1, 1, rep(0, p-3))

X1 <- MASS::mvrnorm(n, mean1, diag(p))
X2 <- MASS::mvrnorm(n, mean2, diag(p))
X3 <- MASS::mvrnorm(n, mean3, diag(p))
X4 <- MASS::mvrnorm(n, mean4, diag(p))

# Combine a "mixture" of 30 + 30 + 20 + 20 = 100 obs from 4 subgroups
X <- rbind(X1[1:30,], X2[1:30,], X3[1:20,], X4[1:20,])
```

(i) Run PCA using `prcomp(X, center = TRUE, scale. = FALSE)`. Report the proportion of variance explained by PC1 and by PC1–PC2 together.

(ii) Create a *scree plot* showing the proportion of variance explained by each principal component.

(d) (10 points) Using the synthetic 25-dimensional dataset from part **(c)**, compare the following two clustering approaches.

(i) Run *k-means* with $k = 4$ on the original 25-dimensional data. Visualize the resulting clusters on the PC1–PC2 plane.

(ii) Project the data onto the first two principal components (PC1 and PC2), then run *k-means* clustering with $k = 4$ on these two-dimensional PC scores.

For each approach, draw a scatter plot on the PC1-PC2 plane, coloring points by the assigned cluster. Then briefly discuss how PCA can help (or might not help) before applying k-means in a high-dimensional setting. Thereafter, for a brief sensitivity check, repeat the PC-score k-means analysis with $k = 3$ and $k = 5$, and comment in one or two sentences.

Problem 2: K-means clustering (35 points total).

You have $n = 10$ data points, each representing a local retailer's attributes. Two features are:

- $x =$ *store floor area* (in hundreds of m^2),
- $y =$ *annual revenue* (in thousands of \$).

Below are the measurements $\{z_i = (x_i, y_i)\}_{i=1}^{10}$:

$$z_1 = (2,\ 4), \quad z_2 = (3,\ 3), \quad z_3 = (3,\ 6), \quad z_4 = (5,\ 15), \quad z_5 = (6,\ 12),$$
$$z_6 = (4,\ 2), \quad z_7 = (10,\ 22), \quad z_8 = (11,\ 20), \quad z_9 = (12,\ 25), \quad z_{10} = (8,\ 16).$$

We want to form $K = 3$ clusters using *k-means* clustering. Use squared Euclidean distances on the raw coordinates throughout this problem.

(a) (10 points) Explain k-means clustering in your own words, including:

- What is the objective function it aims to minimize?
- How does the algorithm alternate between assigning points to clusters and updating centroids?
- Why do we need to pick K in advance, and how might we choose K in practice?

(b) (5 points) Draw a scatter plot of $z_i = (x_i, y_i)$. Briefly describe any obvious grouping you see (if any).

(c) (10 points) Perform two iterations of k-means using $K = 3$ clusters on the data $\{z_1, \ldots, z_{10}\}$ above. Start with an initial assignment:

$$C_1 = \{z_1,\ z_6\}, \quad C_2 = \{z_2,\ z_3,\ z_5\}, \quad C_3 = \{z_4,\ z_7,\ z_8,\ z_9,\ z_{10}\}.$$

Carry out:

(i) Compute centroids of C_1, C_2, C_3.

(ii) Reassign each z_i to the closest centroid.

(iii) Recompute centroids, then reassign again.

Report the centroids used and the resulting memberships for each iteration. Stop after the second reassignment. Show the squared-distance calculations for z_4 in each iteration. You may use R for the remaining arithmetic.

(d) (10 points) Try random initialization: Run k-means with $K = 3$ to convergence from five random initializations. Report the total within-cluster sum of squares for each run and the memberships for a run with the smallest value. If you randomly assign points to 3 clusters at the start, do the five runs give different partitions? Ignore differences that only rename the clusters.

Problem 3: Hierarchical clustering (25 points total).

(a) (5 points) In your own words, explain the hierarchical clustering algorithm, including:

- How does it start, and how are "closest" clusters merged at each step?
- List at least two common linkage rules for measuring dissimilarity between clusters.

(b) (10 points) [JWHT21, Chapter 12, Exercise 4].

(c) (10 points) Consider the dataset from Problem 2. Use hierarchical clustering with **complete linkage** and Euclidean distance on the raw coordinates. Show the successive merges, specifically answering:

- Which points/clusters merge first?
- Is there a noticeable large jump in merge height? If so, identify the largest increase between consecutive merge heights.
- About how many merges occur before a "major" grouping forms?

Sketch or print a dendrogram. Separately, identify a horizontal cut that yields 3 clusters, and compare this partition with the run having the smallest within-cluster sum of squares in Problem 2(d).

Problem 4: Bonus problems (10 bonus points).

(a*) (*5 bonus points) [JWHT21, Chapter 12, Exercise 5].

(b*) (*5 bonus points) [JWHT21, Chapter 12, Exercise 6].

References

[JWHT21] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*, volume 112 of *Springer Texts in Statistics*. Springer, New York, NY, 2nd edition, 2021.