STA 35C – Homework 6

Submission due: Tue, June 3 at 11:59 PM PT

Instructor: Dogyoon Song

Instructions: Upload a PDF file, named with your UC Davis email ID and homework number (e.g., `dgsong_hw6.pdf`) to Canvas ("Homework 6" under "Assignment"). Please make sure to include "STA 35C," your name, and the last four digits of your student ID on the front page. No late homework will be accepted for any reason; submissions after the deadline will receive 0 points. For more details on submission requirements and the late submission policy, see the syllabus.

(**Update on Fri, May 30:** Problem 3 will be graded upon completion.)

(**Update 2 on Sun, June 1:** The data generation code in Problem 1-(c) has been corrected (thanks, Soobin!). With this, you should see more intuitive results in (c) and (d).)

Problem 1: Principal component analysis (40 points total + 5 bonus points).

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 *sample covariance matrix*:

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

(*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 $\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} \begin{bmatrix} b_1 & b_2 \end{bmatrix} = \begin{bmatrix} a_1 b_2 & a_1 b_2 \\ a_2 b_1 & a_2 b_2 \end{bmatrix}$. 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* $\mathbf{u}_1$.

(b) (10 points + 5 bonus)

(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*) **(*5 bonus points)** Verify that these directional variances match $\mathbf{e}_1^\top \, \Sigma \, \mathbf{e}_1$ and $\mathbf{u}_1^\top \, \Sigma \, \mathbf{u}_1$, 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):

```r
set.seed(111)
n  <- 100
p  <- 25

mean1 <- c(-2.5, -2.5, rep(0, p-2))
mean2 <- c(4, rep(0, p-1))
mean3 <- c(2, 3.5, -1, rep(0, p-3))
mean4 <- c(2, 3.3, 1, rep(0, p-3))

# X1 <- matrix(rnorm(n*p, mean=mean1, sd=1.0), n, p)
# X2 <- matrix(rnorm(n*p, mean=mean2, sd=1.0), n, p)
# X3 <- matrix(rnorm(n*p, mean=mean3, sd=1.0), n, p)
# X4 <- matrix(rnorm(n*p, mean=mean4, sd=1.0), n, p)

X1 <- MASS::mvrnorm(n, mean1, diag(1, nrow = p, ncol = p))
X2 <- MASS::mvrnorm(n, mean2, diag(1, nrow = p, ncol = p))
X3 <- MASS::mvrnorm(n, mean3, diag(1, nrow = p, ncol = p))
X4 <- MASS::mvrnorm(n, mean4, diag(1, nrow = p, ncol = 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) Identify the first principal component and compute the directional variance along it.

(ii) Run PCA (e.g. using `prcomp`), and create a *scree plot* showing the variance explained by each principal component.

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

(i) Run *k-means* clustering with $k = 4$ on these 25D points and visualize the results on PC1-PC2 plane.

(ii) Project your data onto the first two principal components (PC1 and PC2), and run *k-means* clustering with $k = 4$ on these 2D projected points.

Draw a scatter plot on the PC1-PC2 plane, coloring points by the assigned cluster, and briefly discuss how PCA can be (or may not be) helpful before doing k-means in a high-dimensional setting. Perform k-means clustering with $k = 3$ and $k = 5$ and comment on your results.

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 $\{(x_i, y_i)\}_{i=1}^{10}$:

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

We want to form $K = 3$ clusters using *k-means* clustering.

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

- 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.

Show your arithmetic through hand calculation or R (stop after 2 full updates).

(d) (10 points) Try random initialization: If you randomly assign points to 3 clusters at the start, does the final solution differ? Repeat k-means with 5 different random starts, and note any differences in the final cluster membership or total within-cluster sum of squares.

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 commonly used distance measures between clusters.

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

(c) (10 points) Consider the dataset from Problem 2. Use hierarchical clustering (e.g. complete-link or average-link). Show the successive merges, specifically answering:

- Which points/clusters merge first?
- About how many merges occur before a "major" grouping forms?

Sketch or print a dendrogram. Identify a horizontal cut that yields 3 clusters, and compare these 3 clusters to the k-means results from Problem 2.

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.