Multivariate Data Analysis. Home Assignment 2

Xijia Liu

Part I: Implement and practice Principal Components Analysis in R. In the next series of exercises, we work on the handwritten digit data set.

Task 1: Read data set into R. Choose your favourite number. Use function 'image' to display the first 24 cases.

Task 2: Do PCA on your data set. Display the first and last 4 principal vectors as 16×16 images.

Task 3: Image reconstruction. Approximate one image in your data set by its first 30, 60, 100, 150, and 200 principal components separately. Put the original image and five approximated images in one plot. For each approximated image, calculate and report the mean square errors, $||\mathbf{x} - \widehat{\mathbf{x}}||^2/p$, where $\mathbf{x}_{p\times 1}$ denotes the original image, $\widehat{\mathbf{x}}_{p\times 1}$ denotes the approximated image, and $p = 256$ is the number of pixels in a image.

Task 4: Calculate how many principal components do you need if you want to keep 85% of information from the original dataset.

Task 5: Build a linear discriminant analysis classifier based on the first two principal components (PCs) to classify the images of digits 5 and 6. To solve this task, the following procedure is suggested.

1. Create a new working dataset that contains all the images of 5 and 6.

2. Split the data set into training set (80%) and testing set (20%) by random sampling.

3. Do PCA on the training set. Use the eigen vectors of the sample covariance matrix of training set to calculate the PCs for both training set and testing set.

4. Use PCs of training set to build a linear discriminant analysis classifier

5. Apply your classifier to the testing set PCs and calculate the accuracy.

Can you achieve such accuracy if only use two original variables to build the classifier? Think about it.

Part II: Theoretical problems. Solve the following theoretical exercises: 8.3 and 8.4 in Applied Multivariate Statistical Analysis (AMSA).

Ex 8.3. Let

$$\mathbf{\Sigma} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix}$$

Determine the principal components $Y_1, Y_2,$ and Y_3. What can you say about the eigenvectors (and principal components) associated with eigenvalues that are not distinct?

Ex 8.4. Calculate the proportion of the total population variance explained by each principal components when the covariance matrix is

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma^2 & \sigma^2\rho & 0 \\ \sigma^2\rho & \sigma^2 & \sigma^2\rho \\ 0 & \sigma^2\rho & \sigma^2 \end{pmatrix}$$

where $-\frac{1}{\sqrt{2}} < \rho < \frac{1}{\sqrt{2}}$.

Please submit your report no later than the 29th of September.