STAT 185 - Homework #2

NAME

This assignment covers multivariate probability and Principal Component Analysis through a mixture of theory and coding. You will need the file *pset2.Rdata* available on Canvas. The commands for loading the data assume this file, *Homework2.Rmd*, and the Rdata file are in the same folder.

Complete the following questions and submit your solutions as a *pdf* via Canvas by Tuesday September 24th at 11:59 PM.

1. Consider a data matrix $X \in \mathbb{R}^{N \times d}$ with centered columns so that the sample covariance

$$\hat{\Sigma} = \frac{X^T X}{N}.$$

Assume $\hat{\Sigma}$ has eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_d > 0$ with orthonormal eigenvectors $\vec{w}_1, \ldots, \vec{w}_d$.

   a. If $X^{(1)} = X - X\vec{w}_1\vec{w}_1^T$ is the data matrix where each row has had its component in the direction of $\vec{w}_1$ removed, show that

$$\hat{\Sigma}^{(1)} = \frac{{X^{(1)}}^T X^{(1)}}{N} = \hat{\Sigma} - \lambda_1 \vec{w}_1 \vec{w}_1^T.$$

   b. Show that $\hat{\Sigma}$ can be written in the form $\hat{\Sigma} = \sum_{j=1}^{d} \lambda_j \vec{w}_j \vec{w}_j^T$.

2. Suppose $\vec{q}_1, \ldots, \vec{q}_k$ are orthonormal vectors in $\mathbb{R}^d$ with $k < d$. Suppose

$$\vec{x} = a_1\vec{q}_1 + \cdots + a_k\vec{q}_k + \vec{\epsilon}$$

where $a_1, \ldots, a_k$ are uncorrelated, mean zero random variables with variance λ^2 and $\vec{\epsilon}$ is independent of $a_1, \ldots, a_k$ and follows a multivariate Normal distribution with mean $\vec{0}$ and covariance $\sigma^2 I$. Here, I is the $d-$dimensional identity matrix.

   a. Find the mean of $\vec{x}$.

   b. Find the covariance matrix, Σ, of $\vec{x}$.

   c. Give the eigenvalues and respective eigenvectors of Σ. (Hint: you may assume there are orthonormal vectors $\vec{q}_{k+1}, \ldots, \vec{q}_d$ which are a basis for the orthogonal complement of the span of $\vec{q}_1, \ldots, \vec{q}_d$.)

   d. Suppose $d = 200$. A sample of 5000 independent realization of $\vec{x}$ were drawn and saved in the 5000×200 data matrix, *problem2*. Compute the principal component analysis of *problem2* using the command princomp.

   e. Generate the scree plot using your results from part d.

   f. Based on your results from parts c and e, estimate k, λ^2, and σ^2. Explain your reasoning.

3. Let $\vec{x}_1, \ldots, \vec{x}_N$ be vectors in $\mathbb{R}^d$. Assume a PCA of these data has loadings $\vec{w}_1, \ldots, \vec{w}_d$ with associated variances $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$. Let U be a $d \times d$ orthonormal matrix and set $\vec{y}_i = U\vec{x}_i$. Find expressions for the principal components loadings and variance of $\vec{y}_1, \ldots, \vec{y}_N$ in terms of U, $\vec{w}_1, \ldots, \vec{w}_d$ and $\lambda_1, \ldots, \lambda_d$.

4. The variable *digits* contains a list of pixelated handwritten digits from the MNIST data set. The pixels are saved in data matrix *digits$pixels*. The labels for each digits are saved in *digits$labels*.

   a. Using princomp, conduct principal component analysis only on the 0 and 1 digits in the dataset.

   b. Generate a scree plot summarizing your results from part a.

c. Describe the geometry of point cloud of images.

d. Plot first two principal scores for each digit with red markers for the zero digits and blue markers for 1 digits. Do you see any clustering in the data? Explain, why or why not.

5. Repeat problem 2, now including all digits in your analysis.

   a. Conduct PCA on all 42000 images.

   b. Generate the scree plot.

   c. Use the command below the plot the digits with the given coloring scheme. Discuss any clustering you do or do not see in the data.

6. Consider a set of points drawn from a 2-dimensional multivariate normal distribution with mean 0 and covariance $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$.

   a. Generate 1000 independent observations from this distribution and generate a scatter plot of these points.

   b. Compute the principal components scores and variances using princomp. Give the 2×2 matrix with the principal component loadings.

   c. Add the point $(10000, 10000)^T$ to these data and rerun the principal component analysis. Give the 2×2 matrix of loadings. Did the loadings change? If so, explain why and how the new loadings are related to the new datum.

7. In this problem, we will use the *mtcar* data set containing 11 observations from 32 cars.

   a. Show the principal component loadings.

   b. Rescale the mpg (miles per gallon) data to feet per gallon, i.e. *mtcars$mpg <- 5280*mtcars$mpg.* Rerun PCA on these modified data and show the loadings.

   c. What is the first loading capturing? Explain this result.

   d. Rerun the results using the empirical correlation matrix by setting the option *cor = TRUE* in the princomp command. Compare this result with part (a)

8. Generate 500 independent realizations from a $10,000$-dimensional normal distribution with mean zero and covariance equal to the identity matrix.

   a. What is the rank of the empirical covariance matrix? How does this compare to the true covariance matrix?

   b. Generate the scree plot for these data by calculating the eigendecomposition of the emprical covariance matrix. Explain the graph and its connection with part (a). (Note: princomp cannot be used since $N < d$.)

   c. Is PCA a good method for reducing the dimensions of these data? Explain.

9. Consider the data in *problem9* which are samples from $\mathbb{R}^3$ which are shown in Figure 1.

   a. What is the dimension of the shape formed by these data? If you generate the graph above in the RStudio console, you can rotate the figure to see more clearly.

   b. Compute the three principal component variances and show them below.

   c. Do the principal component variances reflect the dimension of the data? Why or why not?

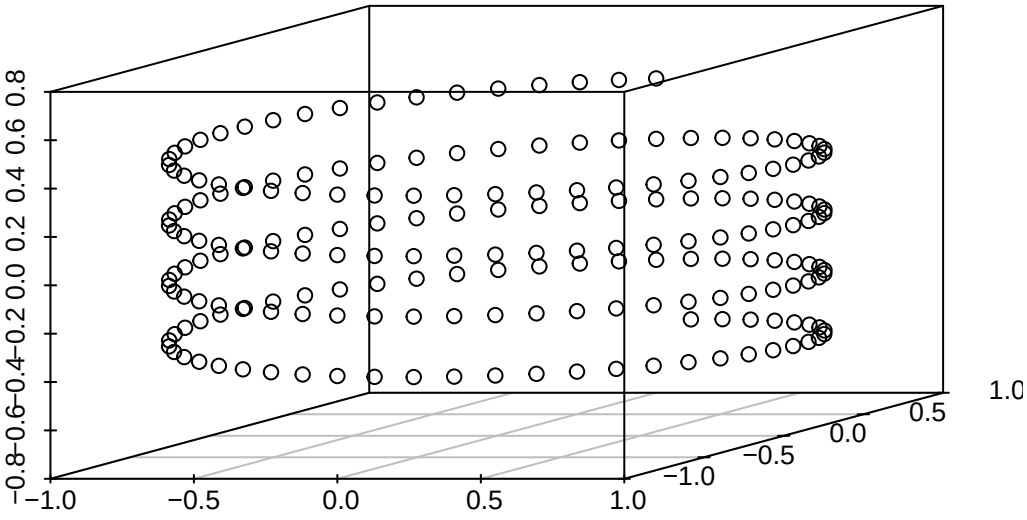

Figure 1: Problem 9 data