Stats 601 Homework 2

Due by 11:59pm on Canvas on Feb 2, 2022

Please scan your solution as a pdf file and submit it to Canvas. Please arrange the pages in order and put your name and uniqname on the top of the first page.

1 (**Gaussian Graphical Model Estimation**) Ex. 17.7 in "Elements of Statistical Learning".

2 (**PCA**) As discussed in class, show that PCA can be viewed as minimizing the approximation error; that is, the optimization problem (following the notation in our lecture note)

$$\min_{\mu, X, U: U^\top U = I_p} \sum_{i=1}^{N} \|Y_i - \mu - U X_i\|^2$$

has solution $\mu = \bar{Y}$, $U = [U_1, \ldots, U_p]$ being the top p eigenvectors of the sample covariance of Y, and $X_i = U^\top (Y_i - \bar{Y})$.

3 (**PCA**) On Canvas site there is an R file `nytimes.RData`. Load this into R workspace, and look for data frame `nyt.frame`. This data frame has 102 rows and 4432 columns. Each row corresponds to a New York Times article. The first column contains the class label of the article (art or music), the remaining columns contains the normalized word counts, one column for each word. You will see a lot of zeroes, because many words do not appear in a given article.

(a) Write your own code for retrieving the principal component directions $\{U_j, j = 1, \ldots, 4431\}$ of the data set.

(b) Project each article onto a subspace with one, two or three dimensions using the principal component directions W_1, W_2 and W_3. Be sure to include the class label (art or music) for each article. How many dimensions do you need to visually separate the two classes of articles reasonably well?

(c) Examining the weights (also called the loadings) of each W_j can give us some insight about the roles of original variables (dimensions). Report the 20 words which correspond to the maximum positive and minimum negative weights for each of the top 3 PC directions. Based on this, can you comment on the role each of the PC direction play in capturing the variation (variance) of the data set (of articles)?

4 **(Factor analysis and PCA)** Generate $n = 100$ observations of the 15-dim vector Y from the following factor analysis model:

$$Y = \mu^* + \Lambda^* X + W$$

where X and W are independent, $\mu^* = \mathbf{0}_{15\times 1}$ and

$$\Lambda^* = \begin{pmatrix} \mathbf{1}_{8\times 1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{7\times 1} \end{pmatrix}_{15\times 2}, X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}_{2\times 1} \sim N_2(0, I_2), \quad W_{15\times 1} \sim N_{15}(0, 0.5\times I_{15}).$$

(a) Show that Λ can only be identified up to an orthornormal transformation. Propose one such suitable distance measure to evaluate the distance of an estimated *column space* of Λ to the true column space (of Λ^*).

(b) Suppose we know that there are two factors in the factor analysis model. Fit the factor analysis model, and what is the equation that defines the estimated subspace of scores X? Plot the projections of the data set onto the 2-dim subspace.

(c) Implement PCA to identify the principal components and the projections of the data set on to the 2-dim principal subspace. Compare your results with those obtained in part (b).

5 **(Estimation of the number of factors)**

(a) [**PCA**] Consider the data generated in Question 4, which can be written as a 100×15 matrix $\mathbf{Y}$. Suppose we don't know the number of PC components and want to use the permutation method to estimate it.

Specifically, for the 100×15 data matrix $\mathbf{Y}$, do random permutation for each column and compute the corresponding eigenvalues of the sample covariance of the permuted data $\tilde{\mathbf{Y}}$.

Repeat the above procedure a large number times, such as 500, and obtain the permutation distributions of the eigenvalues. Use the permutation distributions and the observed eigenvalues from data $\mathbf{Y}$ to decide the number of PC components.

Report your findings.

(b) [**PCA**] Use simulation to check the performance of the permutation procedure in (a). In particular, regenerate data matrix $\mathbf{Y}$ from the factor analysis model in Question 4 many times (such as 100), and for each time, estimate the number of PC components using the method in (a). How many times are the number of components correctly estimated? Report your findings.

(c) [**PCA**] Suppose we select the number of PC components such that the explained proportion of variance is 90%. Use simulation to compare its performance with the permutation method in (b).

(d) [**Factor Analysis**] Consider the data matrix $\mathbf{Y}$ generated in Question 4. Use the likelihood ratio test under the factor analysis model to estimate the number of factors.

(e) [**Factor Analysis**] Use simulation to check the performance of the likelihood ratio test method in (d). Report your findings.