2023 STAT 528: Homework 3

Due Wednesday, February 8th at 11:59pm

1 Finish Updating Writing (20 points)

The goal of this problem is to enhance the scientific and technical writing in your report. In your feedback to Homework 2, we indicated a specific question for you to focus on for Homework 3. Do the following:

- Copy/paste your original text and the question indicated. There is no need to include other sections as well.
- Edit the text of the indicated question based on Zhen's feedback from Homeworks 1 and 2, and the lecture on scientific and technical writing. Below your paragraph, state what improvements you made and how they relate to the lectures on scientific and technical writing. Provide also an answer to the question left by Zhen.

Due to varying quality in the original submissions, we expect some students will have substantial work for this problem while others will only need to make minor changes. Grades will be based on overall quality, not solely improvement.

2 Factor analysis on Psychomotor Test data (25 points)

For this problem, you'll study the Psychomotor Test data. In this data from Fleishman and Hempel (1954), a sample of 197 airmen took a range of tests, divided into three categories. The first category consisted of performance tests of flying ability over eight time trials, with scores given for each trial (variables $x_1, \ldots, x_8$ in the dataset). Then, each airman took written tests to assess speed in verbal, spatial, and arithmetic tasks (x_9, x_{10}, x_{11}) and practical tests in operating various apparati and to asses reaction times and rate of movement (x_{12}, x_{13}, x_{14}). The test data can be downloaded on the page for this assignment. Note that you are only given a sample covariance matrix, and not the original scores. Download the test data and load it into R using the following code:

```
psycho.file<-"psycho.txt"
n<-max(count.fields(psycho.file))
psycho.mat<-data.matrix(read.table(psycho.file,fill=TRUE,col.names=1:n))
colnames(psycho.mat)<-c(paste0("Performance_",c(1:8)),
paste0("Written_",c("Verbal","Spatial","Arithmetic")),
paste0("Practical_",c("Operating","ReactionTime","RateofMovement")))
rownames(psycho.mat)<-colnames(psycho.mat)
```

1. Create and interpret a scree plot to get an idea of how many factors could be used in a factor analysis. Be sure your plot adheres to the principles of good statistical graphics. Then, state your choice for the number of factors to use in an analysis.

2. Regardless of your previous answers, use the R function **factanal** to learn a 2-factor model; plot the unrotated loadings of the factors in a 2-factor model and try to interpret the factors.

3. Perform an Oblimin rotation and plot the loadings of the factors in a 2-factor model. Try to interpret the factors and compare your results to those from the previous question.

3 PCA and VAEs on the MNIST dataset

(50 points) In this section, we explore the methods PCA, Factor analysis, and VAE's for unsupervised learning on the MNIST dataset. The MNIST dataset consists of a large collection of handwritten digits with $60,000$ training images.You may load the dataset from here `https://search.r-project.org/CRAN/refmans/dslabs/html/read_mnist.html`.

3.1 PCA (25 points)

1. Based on the data, decide whether a covariance matrix or correlation matrix is more appropriate to use when forming principal components. Based on your choice, use the R function *princomp* to perform PCA

2. Run the code **names(pca)** to explore outputs of the **princomp** function in R. Then, create a scree plot (being sure it adheres to the principles of good statistical graphics). Do you see an elbow to select the number of principle components?

3. Create a table displaying the percentage of variance and cumulative percentage of variance explained by each principal component. Be sure your table adheres to the principles of good statistical graphics.

4. Regardless of you answers to the previous questions, project all the data into the two principal components (i.e. principle component scores defined in class). Label the points based on the corresponding digits. Do you see any patterns?

3.2 VAEs (25 points)

For this problem, we will use the keras library in R to train a VAE on the MNIST dataset. Please refer to the source `https://github.com/rstudio/keras/blob/main/vignettes/examples/variational_autoencoder.R`.

1. Train a VAE. Use the architecture in the resource provided above. Try different number of latent variables $h = 2, 3, 4$. For each model, take 3 images at random in the training data and visualize the reconstructed images. Compare the reconstruction performance of the three models.

2. Consider the $h = 4$ model. Taking three random images, perform latent traversals. Based on the traversals, what do you think the latent variables represent?

3. Take the $h = 4$ model and consider the two latents with the largest estimated prior variance. Compute the posterior mean of the two latents for all the training images. Plot the two dimensional latents and label the points based on the corresponding digits. Do you see any patterns? As always, adhere to the principle of good statistical graphics.