ORFE 525: Statistical Learning and Nonparametric Estimation
Homework 4

Zachary Hervieux-Moore

Tuesday 18th April, 2017

Exercise 1: In this problem, we will be using K-means clustering on Wikipedia introductions of famous individuals who are associated with Princeton University. Through our analysis, we will determine broad categories that capture the different academic and cultural backgrounds of these individuals. This analysis is part of the field called "topic modeling".

1) Load in the dataset `Wikipedia.RData`. Print out the head of the dataset. Who are the first 3 individuals in the dataset and their corresponding professions? (It's okay to give a broad category). Then use the script `script1_HW4.R` to convert the matrix into a term-document matrix. How many individuals are there? How many words are there? What are the top 10 most frequent words in the entire dataset? List the 0%, 25%, 50%, 75%, and 100% quantiles of the word counts across the entire dataset. The phenomenon you will see can be attributed to "Zipf's Law", saying the distribution of word counts in any text data follows an exponential-like curve. In the later questions, we will like to handle these "over-represented" words.

2) We notice that the top ten words are not very interesting with "the" and "and" being among the most frequent words. One way to combat this is to re-parameterize the matrix. Use the script `script2_HW.R` to use the `tf-idf` parametrization (common in text-analysis). Based on this new `tf-idf` parameterization, what are the 10 words with highest total weight across the entire dataset?

3) Great! Hopefully we are starting to see meeaningful topics appear, but the `tf-idf` parameterization might not be highly interpretable as our `dtm.mat` matrix containing the weights instead of word ounts. We would sitll like to know which words are common in the dataset (for our downstream analysis) but we'll like to remove "common" words. Run the scripts `script3_HW4.R`, which does the following two things: First, it removes words that appear too frequently across all the documents. For example, the word "Princeton" is in every document. Second, it removes the top 300 most frequent words according to Google.

   After running the script, how many words are left in our analysis? What are the top ten words in our remaining `dtm.mat.raw`?

4) We see that the most common words in `dtm.mat.raw` are now similar to the top-weighted words in `dtm.mat`. This is great for our future

interpretability. Now, let's zoom in on one particular person and make sure our data was processed correctly. Search for "Ben Bernanke" in our dataset. Print out the raw text for him according to `dat` (our original dataset). Find in this text the 10 top-weighted words according to `dtm.mat` and the 10 words with the highest word counts according to `dtm.mat.raw`.

5) We will now implement K-means analysis. One of the most important ingredients for K-means is a good distance function. Typically, the Euclidean distance is used between two vectors, but in this question, we will assume that the cosine-distance is more appropriate (`https://en.wikipedia.org/wiki/Cosine_similarity`). We will work with the `dtm.mat` (the `tf-idf` parameterized) matrix when we use K-means (as it includes all the words), but we will use `dtm.mat.raw` to interpret our results (as it includes only the important words we want to see).

   Renormalize all the columns of `dtm.mat` to have norm 1. Then using `norm.sim.ksc` in the `akmeans` package, set the seed to 10 (for reproducibility) and find 8 clusters according to `dtm.mat`. How big are each of the resulting clusters? Based on the cluster assignments, print out the top 25 words in each cluster according to `dtm.mat.raw`. Also, print out the quantiles (0%, 25%, 50%, 75%, 100%) of the word counts with non-zero word counts within each cluster to get a send of how significant the top words are.

Answer: All code for this question is in the appendix below.

1) The first 3 individuals are shown in Table 1.

Table 1: Profession

Individual	Profession
Michel Che	Chemist
Hossein Modarressi	Jurist
Xiao-Gang Wen	Physicist

There are 812 individuals and 6889 words. The top 10 most frequent words are

"the", "and", "univers", "for", "was", "his", "from", "has", "with", "new"

The quantiles are in Table 2.

Table 2: Quantiles for Word Frequency

0%	25%	50%	75%	100%
2	3	5	14	17317

2) Based on the new parametrization, the 10 words with the most weight are

“she”, “her”, “music”, “econom”, “law”, “scienc”, “mathemat”, “new”, “histori”, “research”

3) After running the script, the number of words left is 6633 and the top ten words are

“she”, “her”, “histori”, “econom”, “polit”, “law”, “music”, “mathemat”, “theori”, “physic”

4) The raw text for Ben Bernanke is:

ben shalom bernanke brnki brnangkee born december 13 1953 is an american economist at the brookings institution who served two terms as chairman of the federal reserve the central bank of the united states from 2006 to 2014 during his tenure as chairman bernanke oversaw the federal reserves response to the late2000s financial crisisbefore becoming federal reserve chairman bernanke was a tenured professor at princeton university and chaired the department of economics there from 1996 to september 2002 when he went on public service leavefrom 2002 until 2005 he was a member of the board of governors of the federal reserve system proposed the bernanke doctrine and first discussed the great moderation the theory that traditional business cycles have declined in volatility in recent decades through structural changes that have occurred in the international economy particularly increases in the economic stability of developing nations diminishing the influence of macroeconomic monetary and fiscal policybernanke then served as chairman of president george w bushs council of economic advisers before president bush nominated him to succeed alan greenspan as chairman of the united states federal reserve his first term began february 1

2006 bernanke was confirmed for a second term as chairman on january 28 2010 after being renominated by president barack obama his second term ended february 1 2014 when he was succeeded by janet yellen

The top 10 weighted words in `dtm.mat` in this text are

“bernank”, “reserv”, “chairman”, “feder”, “term”, “bush”, “succeed”, “econom”, “janet”, “volatil”

The highest word counts in `dtm.mat.raw` in this text are

“chairman”, “bernank”, “feder”, “reserv”, “term”, “econom”, “bush”, “februari”, “second”, “succeed”

5) The size of the clusters are in Table 3

Table 3: Size of Clusters

cluster	1	2	3	4	5	6	7	8
size	64	71	55	205	117	52	109	139

The top words by cluster are in Table 4

Table 4: Top Words by Cluster

1	2	3	4
“she”	“physic”	“she”	“she”
“econom”	“theori”	“her”	“polit”
“her”	“mathemat”	“team”	“econom”
“histori”	“she”	“coach”	“law”
“academi”	“her”	“play”	“her”
“theori”	“field”	“band”	“program”
“develop”	“prize”	“season”	“board”
“physic”	“academi”	“their”	“polici”
“comput”	“theoret”	“music”	“affair”
“mathemat”	“quantum”	“assist”	“educ”
“law”	“music”	“record”	“histori”
“advanc”	“contribut”	“head”	“former”
“technolog”	“develop”	“high”	“committe”
“california”	“string”	“jersey”	“social”
“engin”	“known”	“album”	“offic”
“geolog”	“comput”	“citi”	“develop”
“polit”	“philosophi”	“theolog”	“journal”
“press”	“histori”	“histori”	“appoint”
“then”	“then”	“career”	“secur”
“area”	“use”	“design”	“washington”
“had”	“california”	“game”	“foundat”
“join”	“engin”	“had”	“join”
“use”	“mani”	“hockey”	“elect”
“astronomi”	“under”	“began”	“assist”
“career”	“general”	“more”	“dure”

5	6	7	8
“music”	“histori”	“she”	“she”
“she”	“polit”	“her”	“her”
“her”	“her”	“histori”	“law”
“team”	“she”	“econom”	“then”
“play”	“modern”	“literatur”	“board”
“polit”	“philosophi”	“mathemat”	“play”
“perform”	“war”	“review”	“develop”
“histori”	“econom”	“journal”	“human”
“orchestra”	“visit”	“editor”	“career”
“festiv”	“california”	“englisn”	“washington”
“academi”	“european”	“law”	“name”
“record”	“german”	“theolog”	“team”
“econom”	“yale”	“critic”	“use”
“mathemat”	“academi”	“languag”	“foundat”
“theori”	“their”	“teach”	“into”
“compos”	“advanc”	“write”	“mani”
“symphoni”	“press”	“taught”	“were”
“prize”	“germani”	“cultur”	“won”
“london”	“historian”	“theori”	“academi”
“visit”	“physic”	“press”	“citi”
“ensembl”	“recent”	“former”	“had”
“physic”	“under”	“human”	“but”
“law”	“comput”	“music”	“histori”
“press”	“later”	“prize”	“journal”
“program”	“law”	“articl”	“servic”

The quantiles for each cluster are in Table 5

Table 5: Cluster Quantiles

cluster	0%	25%	50%	75%	100%
1	1	1	2	4	80
2	1	1	2	4	110
3	1	1	2	4	89
4	1	1	2	6	223
5	1	1	2	4	234
6	1	1	2	4	51
7	1	1	2	4	144
8	1	1	2	5	110

Code Appendix:

```
# Part 1.1
library(tm)
library(XML)
library(Matrix)
library(slam)
library(SnowballC)
library(akmeans)
library(lsa)

## First three names
dataset <- local(get(load("./Wikipedia.RData")))
dataset[1:3,"name"]

## Script 1
text = dataset[,"text"]
text = iconv(text, to = "utf-8") #some conversion on SMILE needed
corpus = Corpus(VectorSource(text))
dtm.control.raw <- list(tolower = TRUE, removePunctuation = TRUE,
    removeNumbers = TRUE,
                        removestopWords = TRUE, stemming = TRUE, wordLengths
                             = c(3, 15), bounds = list(global = c(2, Inf)))

dtm.raw = DocumentTermMatrix(corpus, control = dtm.control.raw)
dtm.mat.raw = as.matrix(dtm.raw) # our term-document matrix

## Number of individuals and words
print("Number␣of␣individuals:")
dim(dtm.mat.raw)[1]

print("Number␣of␣words:")
dim(dtm.mat.raw)[2]

## 10 most frequent words
words.sum <- colSums(dtm.mat.raw)
words <- order(words.sum, decreasing = T)[1:10]
```

```
print("The␣most␣common␣words␣are:")
colnames(dtm.mat.raw)[words]

## Quantiles
print("The␣quantiles␣are:")
quantile(words.sum, probs = c(0,0.25,0.5,0.75,1))

# Part 1.2

## Script 2
dtm.control <- list(tolower = TRUE, removePunctuation = TRUE,
    removeNumbers = TRUE,
                    removestopWords = TRUE, stemming = TRUE, wordLengths
                        = c(3, 15), bounds = list(global = c(2, Inf)),
                    weighting = function(x){weightTfIdf(x, normalize =
                        FALSE)})

dtm = DocumentTermMatrix(corpus, control = dtm.control)
dtm.mat = as.matrix(dtm)

## 10 most frequent words
words.sum <- colSums(dtm.mat)
words <- order(words.sum, decreasing = T)[1:10]
print("The␣most␣common␣words␣are:")
colnames(dtm.mat)[words]

# Part 1.3

## Script 3
dtm.mat.indicator = dtm.mat.raw
dtm.mat.indicator[dtm.mat.indicator!=0] = 1

word.presence = apply(dtm.mat.indicator, 2, sum)
idx = which(word.presence >= quantile(word.presence, prob = 0.99))
dtm.mat.raw = dtm.mat.raw[,-idx]

common.words = read.csv("google-10000-english.txt", header = F)
idx = which(colnames(dtm.mat) %in% common.words[1:300,1])
dtm.mat.raw = dtm.mat.raw[,-idx]

## How many words left
print("Number␣of␣words␣left:")
dim(dtm.mat.raw)[2]

## 10 most frequent words
words.sum <- colSums(dtm.mat.raw)
words <- order(words.sum, decreasing = T)[1:10]
print("The␣most␣common␣words␣are:")
colnames(dtm.mat.raw)[words]

# Part 1.4

## Original data
ben.id <- which(dataset$name == "Ben␣Bernanke")
dataset[ben.id, "text"]

## Top ten in dtm.mat
```

```
words <- order(dtm.mat[ben.id,], decreasing = T)[1:10]
colnames(dtm.mat)[words]

## Top ten in dtm.mat.raw
words <- order(dtm.mat.raw[ben.id,], decreasing = T)[1:10]
colnames(dtm.mat.raw)[words]

# Part 1.5

## Renormalize
dtm.mat.norm <- t(quick.norm(t(dtm.mat), mod=1))

## Set set and run k means
set.seed(10)
print("Size␣of␣clusters:")
k_means <- norm.sim.ksc(dtm.mat.norm, k=8)
k_means$size

## Top 25 words in each cluster
for (i in 1:8){
  words.sum <- colSums(dtm.mat.raw[which(k_means$cluster == i),])
  words <- order(words.sum, decreasing = T)[1:25]
  print(paste("Cluster␣", i, ":", sep=""))
  print(colnames(dtm.mat.raw)[words])
}

## Quantiles
for (i in 1:8){
  words.sum <- colSums(dtm.mat.raw[which(k_means$cluster == i),])
  words.sum <- words.sum[words.sum > 0]
  print(paste("Cluster␣", i, "␣quantiles:", sep=""))
  print(quantile(words.sum, probs = c(0,0.25,0.5,0.75,1)))
}
```

Exercise 2: In this problem, we are going to use images of 20 Miss Daegu 2013 contestants. The Korean beauty pageant caused some debates on plastic surgery since all the contestants look almost identical. We hope to rigorously analyze the data by PCA and try to visualize the similarities of those contestants. We get the 20 images from the candidates of that competition. The cropped and aligned images can be downloaded from Blackboard. We use function `readPNG` in R package `png` to load the images and get 20 arrays of dimensions $800 \times 580 \times 3$, where 800×580 is the total pixels of each image and 3 denotes RGB's 3 channels, each with a value ranging between 0 and 1. We only consider the face region of row pixel 171 to 400 and column pixel 206 to 375, then stack them together column by column to generate a 20×117300 matrix, and finally apply PCA.

1) **Solving PCA:** Let $X = (X_1, \ldots, X_n)^T$ be the centered n by p data matrix where X_i of length p represents the information of the i^{th} image. Here $n = 20$, $p = 117300$. PCA actually is equivalent to the diagonalization of sample covariance matrix $S = n^{-1}X^TX$. To be specific, left

$$S = \frac{1}{n}X^TX = V\Lambda V^T$$

where $V = (v_1, \ldots, v_p)$ is orthonormal and $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$ is diagonal with non-increasing elements. Let us assume distinct eigenvalues, i.e., $\lambda_i \neq \lambda_j$ for $i \neq j \leq n$. Note, $\lambda_i = 0$ for $i > n$. Calculating the eigendecomposition could be expensive due to the large number of p. We consider instead the decomposition,

$$\frac{1}{n}XX^T = UDU^T$$

where $U = (u_1, \ldots, u_n)$ is orthonormal and $D = \mathrm{diag}(d_1, \ldots, d_n)$ is diagonal with non-increasing elements. Prove $\lambda_i = d_i$ and $v_i = \frac{X^Tu_i}{\sqrt{n\lambda_i}}$ for $i \leq n$. Use this fact in your coding to significantly improve efficiency to computer eigendecompositions.

2) **Graphical Analyses:** Run PCA on the data matrix. Report the following plots and comment on your observations.

- Plot 1: Graph of PCA eigenvalues in decreasing order.
- Plot 2: Visualize the largest 6 eigenfaces by using function `writePNG`. Note that you need to first rescale eigenvectors into the region $[0, 1]$ and create the corresponding array of dimension $230 \times 170 \times 3$.

- Plot 3: Define the coefficients of the i^{th} image onto the k^{th} eigenface as $\eta_{ik} = v_k^T X_i$. Note again X_i is centered. Visualize the 20 images by plotting the coefficients corresponding to the first two eigenfaces.
- Plot 4: Define pairwise similarity of contestants i and j by $s_{ij} = \sum_{k=1}^{6}(\eta_{ik} - \eta_{jk})^2$. For contestant i, by summing up s_{ij} over j, we can get a measure of how the contestant is different from the rest. Plot $\sum_j s_{ij}$ vs contestant number i and identify the 3 most salient contestants.

Answer: All code for this question is in the code appendix below.

1) First we note that

$$\frac{1}{n}XX^T u_i = d_i u_i$$
$$\frac{1}{n}X^T XX^T u_i = d_i X^T u_i$$

Thus, $X^T u_i$ is an eigenvector of $\frac{1}{n}X^T X$ with eigenvlaue d_i. However, we must normalize this eigenvector. We get that the norm is,

$$u_i^T XX^T u_i = n \cdot d_i$$

Thus, we must divide by $\sqrt{nd_i}$. So we conclude that $v_i = \frac{X^T u_i}{\sqrt{nd_i}}$. However, since we now have an orthonormal eigenvector of S, we must have $d_i = \lambda_i$. So we conclude that $v_i = \frac{X^T u_i}{\sqrt{n\lambda_i}}$.

2) The eigenvalue in decreasing order are shown in Figure 1. Values range from 56,600 to close to 10. Note the exponential decrease in size of the eigenvalues. That means that the first eigenvalue is incredibly explanatory of the variance.

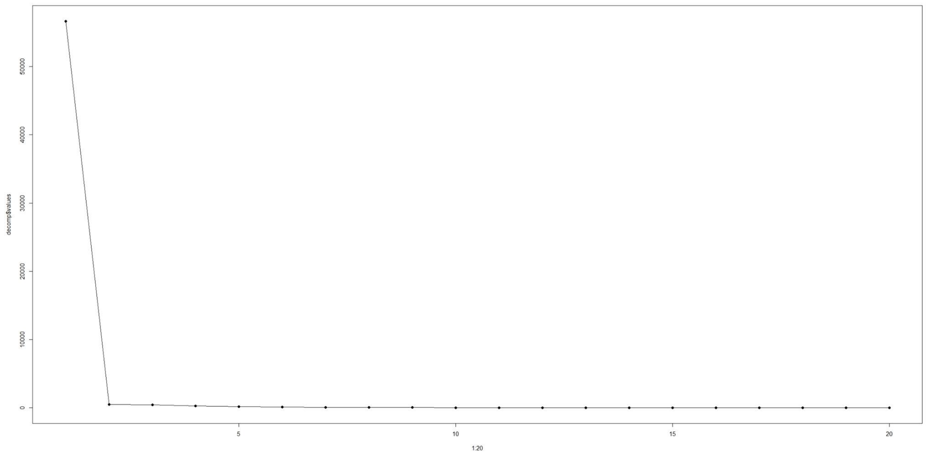
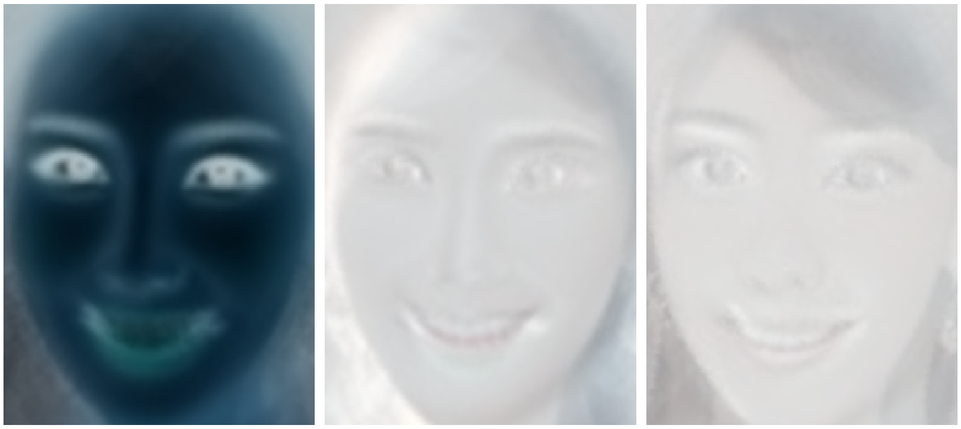


Figure 1: PCA Eigenvalues

The 6 largest eigenfaces are shown in Figure 2.



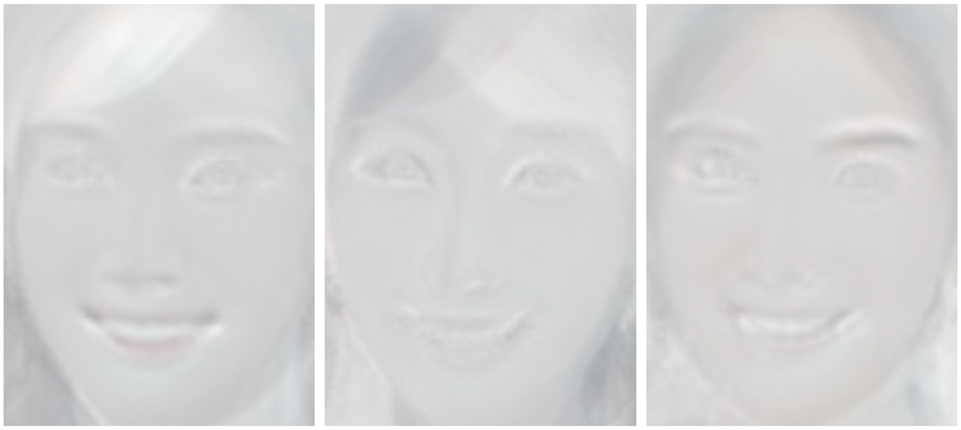


Figure 2: Six Largest Eigenfaces

Figure 3 plots η_{i1}. The values range from -3200 to -4000. This is because the first eigenface is negative. This is obvious by looking at Figure 2 and seeing that it is black. Figure 4 shows η_{i2}. The values range from -700 to 700.

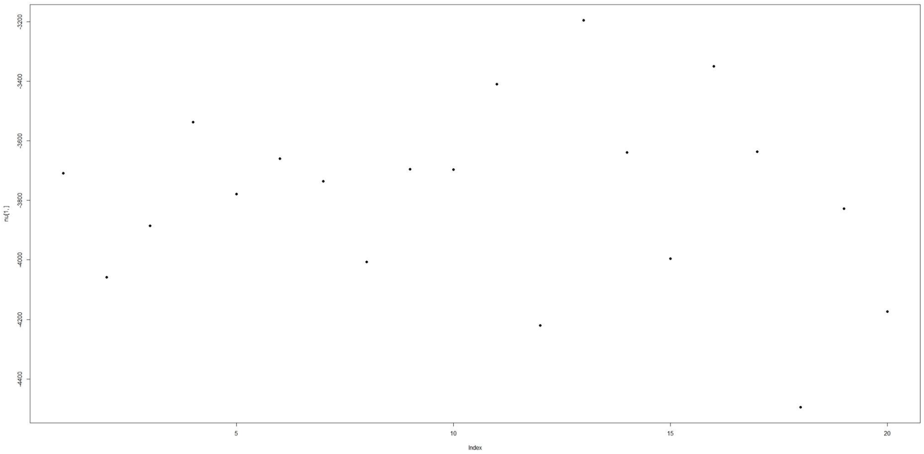


Figure 3: η_{i1} for all Contestants

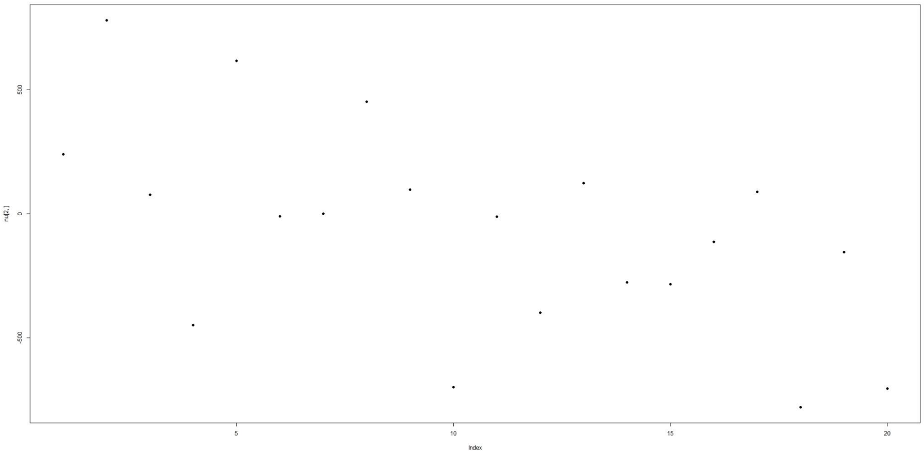


Figure 4: η_{i2} for all Contestants

Figure 5 shows the Plot 4. Values range from $1.5e07$ to $3.5e07$. Note that the vast majority of them are roughly equal in similarity to the rest. Also, they are all the same order of magnitude

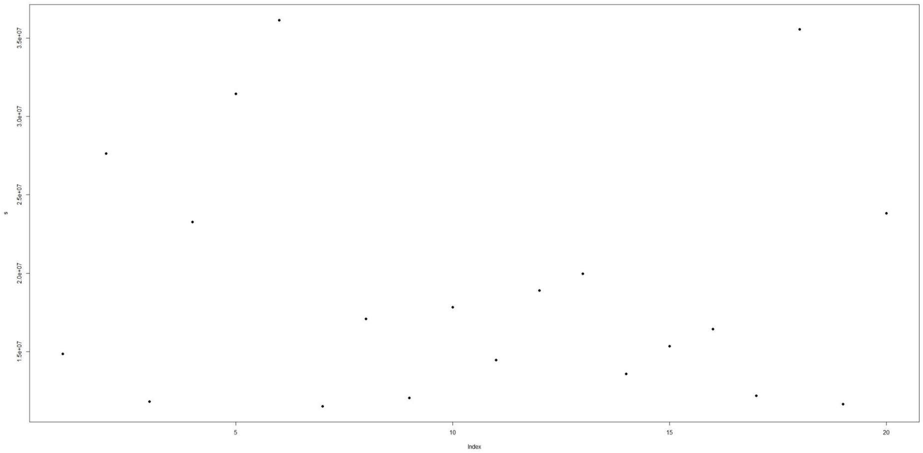


Figure 5: $\sum_j s_{ij}$ for all Contestants

Finally, if salient means contestants that are the most different (highest $\sum_j s_{ij}$) then they are contestants 6, 18, and 5, in that order.

Code Appendix:

```
# Part 2.1
library(png)

n <- 20
p <- 117300

pca <- matrix(0, n, p)

## Pre process images
for(i in 1:20) {
  name <- paste("./images/", i, ".png", sep="")
  p <- readPNG(name)
  p <- p[171:400, 206:375, ]
  pca[i,] <- c(p)
}

## Eigendecomposition using S = XX'/n
decomp <- eigen(pca%*%t(pca)/n)
D <- diag(decomp$values)
U <- decomp$vectors

# Part 2.2

## Plot eigenvalues
dev.new()
plot(1:20, decomp$values, type="o", pch=19)

## Top 6 eigenfaces
V <- t(pca)%*%U
rescale <- function(x){(x-min(x))/(max(x)-min(x))}
V <- rescale(V)

eigenface <- array(rep(0, 230*170*3*6), dim=c(230,170,3,6))

for(i in 1:6) {
  eigenface[1:230,1:170,1,i] = matrix(t(V)[i, 1:39100], 230, 170)
  eigenface[1:230,1:170,2,i] = matrix(t(V)[i, 39101:78200], 230, 170)
  eigenface[1:230,1:170,3,i] = matrix(t(V)[i, 78201:117300], 230, 170)
  writePNG(eigenface[,,,i], target = paste("./eigenface", i, ".png", sep
      =""))
}

## i-th image to k-th eigenface
nu <- matrix(0,6,20)
means <- rowMeans(pca)
pca_center <- pca
## center the pca matrix
for (i in 1:length(means)){
  pca_center[i,] <- pca_center[i,]-means[i]
}

for(i in 1:20) {
  for(k in 1:6) {
    nu[k,i] = t(V)[k,]%*%(pca_center[i,])
  }
}
dev.new()
```

```
plot(nu[1,], pch=19)
dev.new()
plot(nu[2,], pch=19)

## Pairwise similarity
s <- matrix(0,20,20)
for(i in 1:20) {
  for(j in 1:20) {
    s[i,j] = sum((nu[,i] - nu[,j])^2)
  }
}
s <- rowSums(s)
dev.new()
plot(s, pch=19)
print("Most␣salient␣contestants:")
order(s, decreasing=T)[1:3]
```

Exercise 3: In this problem, we will show that K-means algorithm can be derived as a limiting case of the EM algorithm for mixture of spherical Gaussians. That is, for the mixture model $p_\psi(x) = \sum_{j=1}^n p_{\theta_j}(x_j)\eta_j$, where $p_{\theta_j}(x_j)$ is the pdf of $\mathcal{N}(\theta_j, \sigma_j)$, start from the EM algorithm and find th appropriate settings for the E and M steps to reconstruct the K-means algorithm. Justify. (**Hint:** Calculate the limit $\lim_{\sigma_j^2 \to 0} \gamma_{ij}$)

Answer: We have that the mixture for spherical Gaussians is

$$\mathbb{P}_\psi(x) = \sum_{j=1}^n \mathbb{P}_{\theta_j}(x)\eta_j$$

Now, let Y be the random variable of choosing Gaussian j, a categorical distribution, then the joint distribution is

$$\mathbb{P}_\psi(x, y) = \sum_{j=1}^n \mathbb{P}_{\theta_j}(x)1_{\{y=j\}}\eta_j$$

Then, from class, we derived the following E-step for this model

$$\gamma_{ij}^{(t+1)} = \mathbb{P}_{\theta^{(t)}}(Y_i = j | X_i = x_i) = \frac{\mathbb{P}_{\theta_j^{(t)}}(x_i)\eta_j^{(t)}}{\sum_{k=1}^n \mathbb{P}_{\theta_k^{(t)}}(x_i)\eta_k^{(t)}}$$

Now, take $\sigma_j^2 \to 0$, then the above becomes the hard assignment

$$\gamma_{ij} = \begin{cases} 1 & \text{if } \|x_i - \theta_j\|_2^2 < \|x_i - \theta_k\|_2^2, \ \forall\ j \neq k \\ 0 & \text{otherwise} \end{cases}$$

Which is precisely the k-means step of assigning x_i to the cluster with the closest center θ_j. As the σ_j disappear, the E-step only depends on θ_j. Thus, we only need to update the θ_j's. The M-step becomes

$$\theta_j^{(t+1)} = \frac{\sum_{i=1}^n \gamma_{ij}^{(t)} x_i}{\sum_{i=1}^n \gamma_{ij}^{(t)}} = \frac{\sum_{i:\gamma_{ij}^{(t)}=1} x_i}{|\{i : \gamma_{ij}^{(t)} = 1\}|}$$

Thus, the M-step calculates the new center of the cluster to be the average of all the points in the cluster. Thus, the E-step and M-step defined here is precisely the k-means algorithm.

Exercise 4: In this problem, we will derive the EM algorithm for the hidden Markov model (HMM). The HMM is defined as follows. Let $Z_1, \dots, Z_n \in \{1, \dots, k\}$ be a latent discrete Markov chain such that

$$\mathbb{P}(Z_{i+1} = t | Z_i = s) = A_{st}, \text{ for any } i \geq 1$$

and $\mathbb{P}(Z_1 = j) = \eta_j$, where $\sum_{j=1}^k \eta_j = 1$. We observe $X_1, \dots, X_n \in \mathbb{R}^d$ where

$$X_i | Z_i = j \sim \mathcal{N}(\mu_j, \Sigma_j), \text{ for any } i \geq 1$$

Our goal is to estimate $\psi = \{A = [A_{st}]_{s,t=1}^k, \theta_j = (\mu_j, \Sigma_j), \eta = (\eta_1, \dots, \eta_k)^T\}$.

1) Derive the M-step for parameters. Show that the updating rule is

$$\begin{aligned}
\eta_j &\leftarrow \frac{p_{\psi^{\text{old}}}(Z_1 = j | X_1, \dots, X_n)}{\sum_{\ell=1}^k p_{\psi^{\text{old}}}(Z_1 = \ell | X_1, \dots, X_n)} \\
A_{st} &\leftarrow \frac{\sum_{m=2}^n p_{\psi^{\text{old}}}(Z_m = t, Z_{m-1} = s | X_1, \dots, X_n)}{\sum_{\ell=1}^k \sum_{m=2}^n p_{\psi^{\text{old}}}(Z_m = \ell, Z_{m-1} = s | X_1, \dots, X_n)} \\
\mu_j &\leftarrow \frac{\sum_{i=1}^n p_{\psi^{\text{old}}}(Z_i = j | X_1, \dots, X_n) X_i}{\sum_{i=1}^n p_{\psi^{\text{old}}}(Z_i = j | X_1, \dots, X_n)} \\
\Sigma_j &\leftarrow \frac{\sum_{i=1}^n p_{\psi^{\text{old}}}(Z_i = j | X_1, \dots, X_n)(X_i - \bar{X})(X_i - \bar{X})^T}{\sum_{i=1}^n p_{\psi^{\text{old}}}(Z_i = j | X_1, \dots, X_n)}
\end{aligned}$$

2) Derive the E-step to compute the conditional density $p_\psi(Z_i | X_1, \dots, X_n)$ for any $1 \leq j \leq k$. Let $\alpha(Z_i) = p(X_1, \dots, X_n, Z_i)$ and $\beta(Z_i) = p(X_{i+1}, \dots, X_n | Z_i)$. Show that

$$\begin{gathered}
p_\psi(Z_i | X_1, \dots, X_n) \propto \alpha(Z_i)\beta(Z_i) \text{ where} \\
\alpha(Z_i) = p_\psi(X_i | Z_i) \sum_{Z_{i-1}} \alpha(Z_{i-1}) p_\psi(Z_i | Z_{i-1}) \\
\beta(Z_i) = \sum_{Z_{i+1}} \beta(Z_{i+1}) p_\psi(X_{i+1} | Z_{i+1}) p_\psi(Z_{i+1} | Z_i)
\end{gathered}$$

Answer:

1) First, let us write the joint probability

$$\begin{aligned}&\mathbb{P}_\psi(Z_1,\dots,Z_n,X_1,\dots,X_n)\\&=\mathbb{P}_\psi(X_n|Z_1,\dots,Z_n,X_1,\dots,X_{n-1})\cdot\\&\quad\mathbb{P}_\psi(Z_n|Z_1,\dots,Z_{n-1},X_1,\dots,X_{n-1})\cdot\\&\quad\mathbb{P}_\psi(Z_1,\dots,Z_{n-1},X_1,\dots,X_{n-1})\\&=\mathbb{P}_\psi(X_n|Z_n)\cdot\mathbb{P}_\psi(Z_n|Z_{n-1})\cdot\mathbb{P}_\psi(Z_1,\dots,Z_{n-1},X_1,\dots,X_{n-1})\end{aligned}$$

Where we used the Markov property in the last step. Applying the same result recursively yields

$$\mathbb{P}_\psi(Z_1,\dots,Z_n,X_1,\dots,X_n)=\mathbb{P}_\psi(X_1|Z_1)\mathbb{P}_\psi(Z_1)\prod_{i=2}^n\mathbb{P}_\psi(X_i|Z_i)\mathbb{P}_\psi(Z_i|Z_{i-1})$$

Which the above is the definition of the likelihood. Now, we note that

$$X_i|Z_i=j\sim\mathcal{N}(\mu_j,\Sigma_j)$$

$$\implies\mathbb{P}_\psi(X_i|Z_i)=\sum_{j=1}^k\phi_{\mu_j,\Sigma_j}(X_i)1_{\{Z_i=j\}}$$

$$\mathbb{P}_\psi(Z_i|Z_{i-1})=\sum_{j=1}^k\sum_{\ell=1}^k A_{j\ell}1_{\{Z_{i-1}=j\}}1_{\{Z_i=\ell\}}\ \forall\ i\geq 2$$

$$\mathbb{P}_\psi(Z_1)=\sum_{j=1}^k\eta_j1_{\{Z_1=j\}}$$

Thus, the log-likelihood becomes

$$\begin{aligned}\ell_\psi(X,Z)=\sum_{i=2}^n&\left(\sum_{j=1}^k\log\phi_{\mu_j,\Sigma_j}(X_i)1_{\{Z_i=j\}}+\sum_{j=1}^k\sum_{\ell=1}^k\log A_{j\ell}1_{\{Z_{i-1}=j\}}1_{\{Z_i=\ell\}}\right)\\&+\sum_{j=1}^k\log\phi_{\mu_j,\Sigma_j}(X_1)1_{\{Z_1=j\}}+\sum_{j=1}^k\log\eta_j1_{\{Z_1=j\}}\end{aligned}$$

Now, we compute the E-step

$$\mathbb{E}[\ell_\psi(X,Z)|X] = \sum_{i=2}^{n}\Big(\sum_{j=1}^{k}\log(\phi_{\mu_j,\Sigma_j}(X_i))\mathbb{P}_{\psi^{(old)}}(Z_i = j|X_1,\dots,X_n)$$

$$+\sum_{j=1}^{k}\sum_{\ell=1}^{k}\log(A_{j\ell})\mathbb{P}_{\psi^{(old)}}(Z_{i-1}=j, Z_i=\ell|X_1,\dots,X_n)\Big)$$

$$+\sum_{j=1}^{k}\log(\phi_{\mu_j,\Sigma_j}(X_1))\mathbb{P}_{\psi^{(old)}}(Z_1 = j|X_1,\dots,X_n)$$

$$+\sum_{j=1}^{k}\log(\eta_j)\mathbb{P}_{\psi^{(old)}}(Z_1 = j|X_1,\dots,X_n)$$

Now we do the M-step and maximize according to each parameter. Note that we have set up the equation so that we can maximize individual sums as the sums only depend on one set of parameters at a time. Let's start with η_j. We have the constraint that $\sum_{j=1}^{k}\eta_j = 1$. Using a Lagrange multipler, the first order condition is

$$\frac{\mathbb{P}_{\psi^{(old)}}(Z_1 = j|X_1,\dots,X_n)}{\eta_j} = \lambda, \quad \forall\ j$$

By our constraint, $\lambda = \sum_{j=1}^{k}\mathbb{P}_{\psi^{(old)}}(Z_1 = j|X_1,\dots,X_n)$, thus

$$\eta_j \leftarrow \frac{\mathbb{P}_{\psi^{\text{old}}}(Z_1 = j|X_1,\dots,X_n)}{\sum_{\ell=1}^{k}\mathbb{P}_{\psi^{\text{old}}}(Z_1 = \ell|X_1,\dots,X_n)}$$

Similarly, for $A_{j\ell}$, we get the first order condition

$$\frac{\sum_{i=2}^{n}\mathbb{P}_{\psi^{(old)}}(Z_{i-1}=j, Z_i=\ell|X_1,\dots,X_n)}{A_{j\ell}} = \lambda_j, \quad \forall\ j,\ell$$

Again, due to transition probabilities summing up to one,

$$\lambda_j = \sum_{m=1}^{k}\sum_{i=2}^{n}\mathbb{P}_{\psi^{(old)}}(Z_{i-1}=j, Z_i=m|X_1,\dots,X_n)$$

Thus, the update rule is

$$A_{st} \leftarrow \frac{\sum_{m=2}^{n} \mathbb{P}_{\psi^{\text{old}}}(Z_m = t, Z_{m-1} = s|X_1, \dots, X_n)}{\sum_{\ell=1}^{k} \sum_{m=2}^{n} \mathbb{P}_{\psi^{\text{old}}}(Z_m = \ell, Z_{m-1} = s|X_1, \dots, X_n)}$$

Now, we get to the parameters μ_j and Σ_j. We have that we want to maximize the following

$$\sum_{i=1}^{n} \sum_{j=1}^{k} \log(\phi_{\mu_j,\Sigma_j}(X_i))\mathbb{P}_{\psi^{(old)}}(Z_i = j|X_1, \dots, X_n)$$

However, we note that this is the same thing we maximize for the Gaussian mixture model updates which we did in class. Thus, we have the two last update rules

$$\begin{aligned}
\mu_j &\leftarrow \frac{\sum_{i=1}^{n} \mathbb{P}_{\psi^{\text{old}}}(Z_i = j|X_1, \dots, X_n)X_i}{\sum_{i=1}^{n} \mathbb{P}_{\psi^{\text{old}}}(Z_i = j|X_1, \dots, X_n)} \\
\Sigma_j &\leftarrow \frac{\sum_{i=1}^{n} \mathbb{P}_{\psi^{\text{old}}}(Z_i = j|X_1, \dots, X_n)(X_i - \bar{X})(X_i - \bar{X})^T}{\sum_{i=1}^{n} \mathbb{P}_{\psi^{\text{old}}}(Z_i = j|X_1, \dots, X_n)}
\end{aligned}$$

2) The first relation results from

$$\begin{aligned}
\mathbb{P}_{\psi}(Z_i|X_1, \dots, X_n) &= \frac{\mathbb{P}_{\psi}(Z_i, X_1, \dots, X_n)}{\mathbb{P}_{\psi}(X_1, \dots, X_n)} \\
&\propto \mathbb{P}_{\psi}(Z_i, X_1, \dots, X_n) \\
&\propto \mathbb{P}_{\psi}(X_{i+1}, \dots, X_n|Z_i, X_1, \dots, X_i)\mathbb{P}_{\psi}(Z_i, X_1, \dots, X_i) \\
&\propto \mathbb{P}_{\psi}(X_{i+1}, \dots, X_n|Z_i)\mathbb{P}_{\psi}(Z_i, X_1, \dots, X_i) \\
&\propto \beta(Z_i)\alpha(Z_i)
\end{aligned}$$

Now, we show the recursive relation of $\alpha(Z_i)$

$$\begin{aligned}
\alpha(Z_i) &= \mathbb{P}_{\psi}(Z_i, X_1, \dots, X_i) \\
&= \sum_{j=1}^{k} \mathbb{P}_{\psi}(Z_{i-1} = j, Z_i, X_1, \dots, X_i) \\
&= \sum_{j=1}^{k} \mathbb{P}_{\psi}(X_i|Z_{i-1} = j, Z_i, X_1, \dots, X_{i-1})\mathbb{P}_{\psi}(Z_{i-1} = j, Z_i, X_1, \dots, X_{i-1})
\end{aligned}$$

Using the Markov property

$$
\begin{aligned}
&= \sum_{j=1}^{k} \mathbb{P}_\psi(X_i|Z_i)\mathbb{P}_\psi(Z_{i-1}=j, Z_i, X_1, \ldots, X_{i-1}) \\
&= \mathbb{P}_\psi(X_i|Z_i) \sum_{j=1}^{k} \mathbb{P}_\psi(Z_i|Z_{i-1}=j, X_1, \ldots, X_{i-1})\mathbb{P}_\psi(Z_{i-1}=j, X_1, \ldots, X_{i-1}) \\
&= \mathbb{P}_\psi(X_i|Z_i) \sum_{j=1}^{k} \mathbb{P}_\psi(Z_i|Z_{i-1}=j)\alpha(Z_{i-1}=j) \\
&= \mathbb{P}_\psi(X_i|Z_i) \sum_{Z_{i-1}} \mathbb{P}_\psi(Z_i|Z_{i-1})\alpha(Z_{i-1})
\end{aligned}
$$

Now, for $\beta(Z_i)$,

$$
\begin{aligned}
\beta(Z_i) &= \mathbb{P}_\psi(X_{i+1}, \ldots, X_n|Z_i) \\
&= \sum_{j=1}^{k} \mathbb{P}_\psi(X_{i+1}, \ldots, X_n, Z_{i+1}=j|Z_i) \\
&= \sum_{j=1}^{k} \mathbb{P}_\psi(X_{i+1}, \ldots, X_n|Z_i, Z_{i+1}=j)\mathbb{P}_\psi(Z_{i+1}=j|Z_i)
\end{aligned}
$$

Using the Markov property

$$
\begin{aligned}
&= \sum_{j=1}^{k} \mathbb{P}_\psi(X_{i+1}, \ldots, X_n|Z_{i+1}=j)\mathbb{P}_\psi(Z_{i+1}=j|Z_i) \\
&= \sum_{j=1}^{k} \Big(\mathbb{P}_\psi(X_{i+2}, \ldots, X_n|Z_{i+1}=j, X_{i+1})\mathbb{P}_\psi(X_{i+1}|Z_{i+1}=j) \\
&\qquad \mathbb{P}_\psi(Z_{i+1}=j|Z_i)\Big) \\
&= \sum_{j=1}^{k} \beta(Z_{i+1}=j)\mathbb{P}_\psi(X_{i+1}|Z_{i+1}=j)\mathbb{P}_\psi(Z_{i+1}=j|Z_i) \\
&= \sum_{Z_{i+1}} \beta(Z_{i+1})\mathbb{P}_\psi(X_{i+1}|Z_{i+1})\mathbb{P}_\psi(Z_{i+1}|Z_i)
\end{aligned}
$$

Exercise 5: In this problem, we study the EM algorithm for the probabilistic PCA model. We consider the model $X = WZ + \epsilon$, where $Z \sim \mathcal{N}(0, I_q)$ is the latent variable and $\epsilon \sim \mathcal{N}(0, \sigma^2 I_d)$ is independent of Z. The matrix $W \in \mathbb{R}^{d\times q}$ is the linear transform matrix, where $q < d$. Suppose we observe i.i.d. data $X_1, \ldots, X_n$ from this model, where $d < n$. Our goal is to estimate W and σ^2 by EM algorithm.

1) What is the conditional distribution of $Z|X$?

2) (Identifiability condition) Suppose $X = WZ + \epsilon$ and $\tilde{X} = VZ + \epsilon'$ both follow the probabilistic PCA model for $\epsilon \sim \mathcal{N}(0, \sigma^2 I_d)$ and $\epsilon \sim \mathcal{N}(0, \sigma'^2 I_d)$. If $X \overset{d}{=} \tilde{X}$, prove that $\sigma = \sigma'$ and there exists an orthogonal matrix $Q \in \mathbb{R}^{q\times q}$ such that $W = VQ$.

3) Suppose σ^2 is known and the minimum eigenvalue of the sample covariance of $\{X_i\}_{i=1}^n$ is larger than σ^2. What is the maximum likelihood estimator of W? (**Hint 1:** Consider the stationary points of the log-likelihood of the model first. **Hint 2:** Plug the singular decomposition of W into the equation derived by Hint 1 and prove the general formulation of MLE of W. **Hint 3:** By the indentifiability in the previous question, the MLE of W should be defined up to an orthogonal matrix.)

4) Suppose the minimum eigenvalue of the sample covariance of $\{X_i\}_{i=1}^n$ is larger than σ^2. What is the maximum likelihood estimator of W and σ^2? (**Hint:** Use the result in the previous problem.)

5) Write down the EM algoirthm to estimate both W and σ^2. Show your work to get full credit.

Answer:

1) Let us treat (X, Z) as a joint Gaussian vector. X is Gaussian as it is a linear combination of Gaussians. Now, we can treat this pair as being distributed as

$$(X, Z) \sim \mathcal{N}\left(\begin{bmatrix}\mu_X \\ \mu_Z\end{bmatrix}, \begin{bmatrix}\Sigma_{XX} & \Sigma_{XZ} \\ \Sigma_{ZX} & \Sigma_{ZZ}\end{bmatrix}\right)$$

We know that $\mu_Z = 0$ and this implies that $\mu_X = 0$ as ϵ has 0 mean. Also given is $\Sigma_{ZZ} = I_q$. Now, we compute Σ_{XX}.

$$\begin{aligned}\Sigma_{XX} &= \mathbb{E}[XX^T] = \mathbb{E}[(WZ+\epsilon)(WZ+\epsilon)^T] \\ &= \mathbb{E}[WZZ^TW^T + WZ\epsilon^T + \epsilon Z^TW^T + \epsilon\epsilon^T]\end{aligned}$$

Using the fact that $\epsilon \perp\!\!\!\perp X$, the two middle terms are 0,

$$\begin{aligned}
&= W\mathbb{E}[ZZ^T]W^T + \mathbb{E}[\epsilon\epsilon^T] \\
&= WI_qW^T + \sigma^2 I_d \\
&= WW^T + \sigma^2 I_d
\end{aligned}$$

The covariance term is

$$\Sigma_{XZ} = \mathbb{E}[XZ^T] = \mathbb{E}[(WZ+\epsilon)Z^T] \overset{X\perp\!\!\!\perp\epsilon}{=} \mathbb{E}[WZZ^T] = W$$

Lastly, $\Sigma_{ZX} = \Sigma_{XZ}^T$. Now we know all the means and covariances, we can use the formula for conditional Gaussians which yields

$$Z|X \sim \mathcal{N}(W^T(WW^T+\sigma^2 I_d)^{-1}X, I_q - W^T(WW^T+\sigma^2 I_d)^{-1}W)$$

2) As we have $X \overset{d}{=} \tilde{X}$, then their variances must be equal. Then we must have

$$WW^T + \sigma^2 I_d = VV^T + \sigma'^2 I_d$$

Now, since WW^T and VV^T are symmetric, they have diagonalizations

$$\begin{aligned}
WW^T &= P\Lambda_W P^T \\
VV^T &= P'\Lambda_V P'^T
\end{aligned}$$

Thus, we have

$$P(\Lambda_W + \sigma^2 I_d)P^T = P'(\Lambda_V + \sigma'^2 I_d)P'^T$$

Since, $q < d$, the nullspace of W and V is not empty. Thus, Λ_W and Λ_V have a 0 eigenvalue. This means that the minimal eigenvalue for $\Lambda_W + \sigma^2 I_d$ and $\Lambda_V + \sigma'^2 I_d$ are σ and σ' respectively. Thus we must have $\sigma = \sigma'$. Furthermore, by the same reasoning, with $\sigma = \sigma'$, all eigenvalues must be the same. If all the eigenvalue must be the same, then the equation above dictates that $P = P'$ and so they have the same eigendecomposition. Thus, by their SVD, we have

$$\begin{aligned}
W &= P\Sigma_W U \\
V &= P\Sigma_V U'
\end{aligned}$$

Where P is the same in both as their decomposition from above is identical. We must also have $\Sigma_W = \Sigma_V$ from above. Both U and U' are orthogonal as it is a SVD and so there exists orthogonal Q with $U = U'Q$. So we conclude that the same Q also yields $W = VQ$.

3) Recall that $X \sim \mathcal{N}(0, WW^T + \sigma^2 I_d)$. Then we have that the log-likelihood is

$$\ell(X) = \sum_{i=1}^{n} \log \phi_{0,WW^T+\sigma^2 I_d}(X_i)$$

$$= -\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log|WW^T + \sigma^2 I_d| - \frac{n}{2}\sum_{i=1}^{n} X_i^T (WW^T + \sigma^2 I_d)^{-1} X_i$$

$$= -\frac{nd}{2}\log(2\pi) - \frac{n}{2}\log|WW^T + \sigma^2 I_d|$$

$$-\frac{n}{2}\mathrm{Tr}\left(\sum_{i=1}^{n}(X_i X_i^T)(WW^T + \sigma^2 I_d)^{-1}\right)$$

Now, letting $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(X_i X_i^T$, taking the matrix differentiation of the above yields, after a long and arduous process,

$$\frac{\partial \ell}{\partial W} = n((WW^T + \sigma^2 I_d)^{-1}\hat{\Sigma}(WW^T + \sigma^2 I_d)^{-1} W - (WW^T + \sigma^2 I_d)^{-1} W)$$

Which means the stationary points are

$$\hat{\Sigma}(WW^T + \sigma^2 I_d)^{-1} W = W$$

We ignore the trivial solution $W = 0$. Following the hint, let $W = U\Sigma V^T$ be the SVD decomposition. We plug this into the above.

$$\hat{\Sigma}(U\Sigma^2 U^T + \sigma^2 I_d)^{-1} U\Sigma V^T = U\Sigma V^T$$

Post multiplying by V, and since the inverted term is diagonal, and $U\Sigma^2 U^T = \Sigma^2$,

$$\hat{\Sigma} U\Sigma = U(U\Sigma^2 U^T + \sigma^2 I_d)\Sigma$$

Now, take an eigenvector of U, says u_i, such that $\sigma_i \neq 0$ (one of the elements of Σ), then

$$\hat{\Sigma} u_i = (\sigma_i^2 + \sigma^2) u_i$$

So u_i is an eigenvector of $\hat{\Sigma}$ with eigenvalue $\lambda_i = \sigma_i^2 + \sigma^2$. Thus, the stationary points of the MLE are

$$W = U\Sigma V^T$$

Where U is picked such that the first r columns are eigenvectors of $\hat{\Sigma}$, the rest are arbitrary. Plugging this back into the log-likelihood, using the fact that the determinant is the product of eigenvalues, and the trace is the sum of eigenvalues,

$$\begin{aligned}
\ell(X) &\propto -\log|WW^T + \sigma^2 I_d| - \operatorname{Tr}\left(\hat{\Sigma}(WW^T + \sigma^2 I_d)^{-1}\right) \\
&= -\log|U\Sigma^2 U^T + \sigma^2 I_d| - \operatorname{Tr}\left(\hat{\Sigma}(U\Sigma^2 U^T + \sigma^2 I_d)^{-1}\right) \\
&= -\sum_{i=1}^{r} \log(\lambda_i) - \sum_{i=r+1}^{d} \log(\sigma^2) - \sum_{i=1}^{r} 1 - \sum_{i=r+1}^{d} \frac{\lambda_i}{\sigma^2}
\end{aligned}$$

Where the trace simplifies to the above because $\hat{\Sigma} = U\text{diag}(\lambda_1 + \ldots, \lambda_r, \sigma^2, \ldots, \sigma^2)U^T$. Therefore, we must pick a suitable r that maximizes this. Noting that $-\log(\lambda_i) + \log(\sigma^2) \geq 1 - \frac{\lambda_i}{\sigma^2}$ from log inequalities, we want to maximize r to maximize likelihood, thus $r = q$. Then pick W to be the SVD with U having q eigenvector columns of $\hat{\Sigma}$.

4) The MLE for W does not change from the last part. From the previous part, we get that the log likelihood is

$$\ell(X) \propto -\sum_{i=1}^{q} \log(\lambda_i) - \sum_{i=q+1}^{d} \log(\sigma^2) - \sum_{i=1}^{q} 1 - \sum_{i=q+1}^{d} \frac{\lambda_i}{\sigma^2}$$

The first order condition for this is

$$\begin{aligned}
\frac{\partial \ell}{\partial \sigma^2} &= -\sum_{i=q+1}^{d} \frac{1}{\sigma^2} + \sum_{i=q+1}^{d} \frac{\lambda_i}{\sigma^4} = 0 \\
\implies \sigma^2 &= \frac{1}{d-q} \sum_{i=q+1}^{d} \lambda_i
\end{aligned}$$

Which is the MLE for σ.

5) Recall the joint density of $(X, Z) \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 I_d + WW^T & W^T \\ W & I_q \end{bmatrix}\right)$. Now, the log-likelihodd of a Gaussian is

$$\ell_\psi(X, Z) \propto -n\log|\Sigma| - \sum_{i=1}^{n} \begin{bmatrix} X_i^T & Z_i^T \end{bmatrix} \Sigma^{-1} \begin{bmatrix} X_i \\ Z_i \end{bmatrix}$$

Where $\Sigma = \begin{bmatrix} \sigma^2 I_d + WW^T & W^T \\ W & I_q \end{bmatrix}$. Now note that

$$\Sigma = \begin{bmatrix} I_d & W^T \\ 0 & I_q \end{bmatrix} \begin{bmatrix} \sigma^2 I_d & 0 \\ 0 & I_q \end{bmatrix} \begin{bmatrix} I_d & 0 \\ W & I_q \end{bmatrix}$$

Thus,

$$\begin{aligned} \det \Sigma &= \det \begin{bmatrix} I_d & W^T \\ 0 & I_q \end{bmatrix} \cdot \det \begin{bmatrix} \sigma^2 I_d & 0 \\ 0 & I_q \end{bmatrix} \cdot \det \begin{bmatrix} I_d & 0 \\ W & I_q \end{bmatrix} \\ &= 1 \cdot \sigma^{2^d} \cdot 1 \\ &= \sigma^{2^d} \end{aligned}$$

Deriving the inverse from the above is simple, we just find the inverse for the simpler matrices and reverse the order,

$$\Sigma^{-1} = \begin{bmatrix} I_d & 0 \\ -W & I_q \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma^2} I_d & 0 \\ 0 & I_q \end{bmatrix} \begin{bmatrix} I_d & -W^T \\ 0 & I_q \end{bmatrix}$$

This makes the log-likelihood become

$$\begin{aligned} \ell_\psi(X, Z) \propto &- nd \log \sigma^2 \\ &- \sum_{i=1}^{n} \begin{bmatrix} X_i^T & Z_i^T \end{bmatrix} \begin{bmatrix} I_d & 0 \\ -W & I_q \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma^2} I_d & 0 \\ 0 & I_q \end{bmatrix} \begin{bmatrix} I_d & -W^T \\ 0 & I_q \end{bmatrix} \begin{bmatrix} X_i \\ Z_i \end{bmatrix} \\ &= -nd \log \sigma^2 - \sum_{i=1}^{n} \begin{bmatrix} X_i^T - Z_i^T W & Z_i^T \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma^2} I_d & 0 \\ 0 & I_q \end{bmatrix} \begin{bmatrix} X_i - W^T Z_i \\ Z_i \end{bmatrix} \\ &= -nd \log \sigma^2 \\ &+ \sum_{i=1}^{n} \frac{1}{\sigma^2} (X_i^T X_i - 2 X_i^T W Z_i + Z_i^T W^T W Z_i) + Z_i^T Z_i \end{aligned}$$

To do the E-step, we take conditional expectation on X

$$\begin{aligned}\mathbb{E}[\ell_\psi(X,Z)|X] = &- nd\log\sigma^2 - \sum_{i=1}^n \frac{X_i^TX_i}{\sigma^2} \\ &+ 2\sum_{i=1}^n \frac{\mathrm{Tr}(X_i^TW\mathbb{E}_{\psi^{(\mathrm{old})}}[Z_i|X_i])}{\sigma^2} \\ &- \sum_{i=1}^n \frac{\mathrm{Tr}(W\mathbb{E}_{\psi^{(\mathrm{old})}}[Z_iZ_i^T|X_i]W^T)}{\sigma^2} \\ &- \sum_{i=1}^n \mathbb{E}_{\psi^{(\mathrm{old})}}[Z_i^TZ_i|X_i]\end{aligned}$$

Where the conditional expectations can be explicitly solved for using part 1). Now, we do the M-step. We use the fact that $\frac{\partial}{\partial X}\mathrm{Tr}(AXB) = BA$ and $\frac{\partial}{\partial X}\mathrm{Tr}(XBX^T) = XB^T + XB$. Thus, taking the derivative of the log-likelihood with respect to W and setting to 0 yields

$$\sum_{i=1}^n 2X_i\mathbb{E}_{\psi^{(\mathrm{old})}}[Z_i|X_i]^T - W\mathbb{E}_{\psi^{(\mathrm{old})}}[Z_iZ_i^T|X_i] + W\mathbb{E}_{\psi^{(\mathrm{old})}}[Z_iZ_i^T|X_i]^T = 0$$

But, $\mathbb{E}_{\psi^{(\mathrm{old})}}[Z_iZ_i^T|X_i]$ is symmetric, so

$$\begin{aligned}\sum_{i=1}^n X_i\mathbb{E}_{\psi^{(\mathrm{old})}}[Z_i|X_i]^T &= W\sum_{i=1}^n \mathbb{E}_{\psi^{(\mathrm{old})}}[Z_iZ_i^T|X_i] \\ \implies W \leftarrow \left(\sum_{i=1}^n X_i\mathbb{E}_{\psi^{(\mathrm{old})}}[Z_i|X_i]^T\right) &\left(\sum_{i=1}^n \mathbb{E}_{\psi^{(\mathrm{old})}}[Z_iZ_i^T|X_i]\right)^{-1}\end{aligned}$$

Finally, estimate σ^2 by maximizing the E-step w.r.t. σ, first order conditions are

$$
\begin{aligned}
0 =& -\frac{nd}{\sigma^2} + \frac{1}{\sigma^4}\sum_{i=1}^{n} X_i^T X_i \\
& - \frac{2}{\sigma^4}\mathrm{Tr}(X_i^T W \mathbb{E}_{\psi^{(\mathrm{old})}}[Z_i|X_i]) + \frac{1}{\sigma^4}\sum_{i=1}^{n}\mathrm{Tr}(W\mathbb{E}_{\psi^{(\mathrm{old})}}[Z_i Z_i^T|X_i]W^T) \\
\implies \sigma^2 \leftarrow & \frac{1}{nd}\sum_{i=1}^{n} X_i^T X_i \\
& - \frac{2}{nd}\mathrm{Tr}(X_i^T W \mathbb{E}_{\psi^{(\mathrm{old})}}[Z_i|X_i]) + \frac{1}{nd}\sum_{i=1}^{n}\mathrm{Tr}(W\mathbb{E}_{\psi^{(\mathrm{old})}}[Z_i Z_i^T|X_i]W^T)
\end{aligned}
$$