Math 189: Homework 7

Alan Lui, Derek So, Xiangyu Wei

March 7, 2021

Introduction

Nutrition is a complicated matter of balancing dozens if not hundreds of chemical compounds in the human body, along with the variance that exists between each unique individual's needs. Thus, developing a factor model to pinpoint the main drivers of nutrition and reduce the dimensionality of such a problem, as well as discovering the latent patterns of nutrition, is ideal in studying the subject. This study looks into the recommended women's nutritional intake, sampling from data gathered in the United States Department of Agriculture (USDA) Women's Health Survey dataset (nutrient.txt). Using a 1985 USDA Women's Health dataset, we analyze 5 variables from a sample of 737 and search for latent factors that can explain the major nutritional features.

Tasks & Analysis

The following is the packages that are needed in this report

```
library(lattice)
library(ellipse)
```

Dataset

The dataset is from a study of women's nutrition commissioned by USDA in 1985. Nutrient intake of five nutritional components (Calcium, Iron, Protein, Vitamin A and Vitamin C) were collected from a random sample of 737 women aged 25-50 years. The units of these variables are measured in different magnitudes of grams. Calcium, Iron, and Vitamin C are measured in milligrams, Protein in grams, and Vitamin A in micrograms. The dataset is downloaded from https://github.com/tuckermcelroy/ma189/blob/main/Data/nutrient.txt at 2021-01-28 20:15:24 PST.

```
nutrient <- read.table("Data/nutrient.txt")[,-1] # load data
col.names <- c("Calcium", "Iron", "Protein",
               "Vitamin A", "Vitamin C") # new column names
colnames(nutrient) <-  col.names # rename column
head(nutrient, 5)
```

```
##   Calcium   Iron Protein Vitamin A Vitamin C
## 1  522.29 10.188  42.561    349.13    54.141
## 2  343.32  4.113  67.793    266.99    24.839
## 3  858.26 13.741  59.933    667.90   155.455
## 4  575.98 13.245  42.215    792.23   224.688
## 5 1927.50 18.919 111.316    740.27    80.961
```

Methods

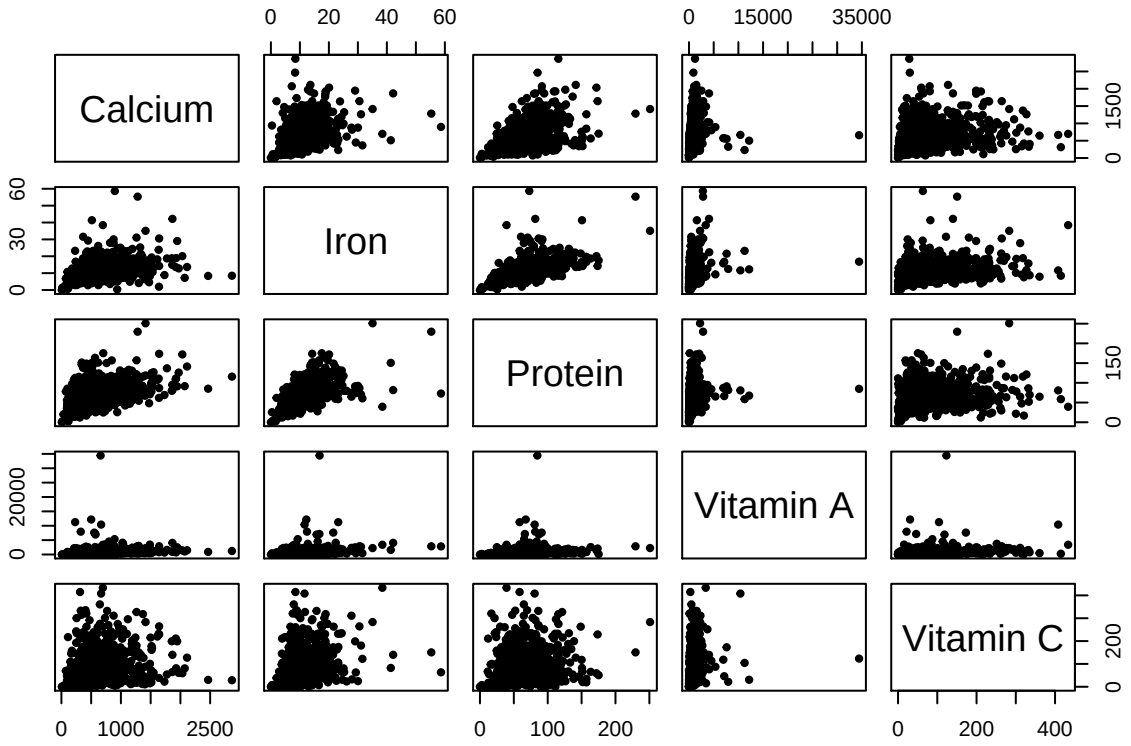
After a level plot that shows the correlation between the five different variables, we use principal component analysis (PCA) and maximum likelihood estimation (MLE) for factor analysis to investigate the factor model that explain the main drivers of nutrition. We will compare the two methods, PCA and MLE, through the proportion of variance explained by each component/factor using a scree plot, the loadings of each variable on the component/factor, and the factor scores using a scatter plot.

Analysis

1. Explore the data graphically in order to investigate the correlations between variables.

The following is a pairwise scatterplot

```
pairs(nutrient, pch = 20)
```



The following is the correlation matrix

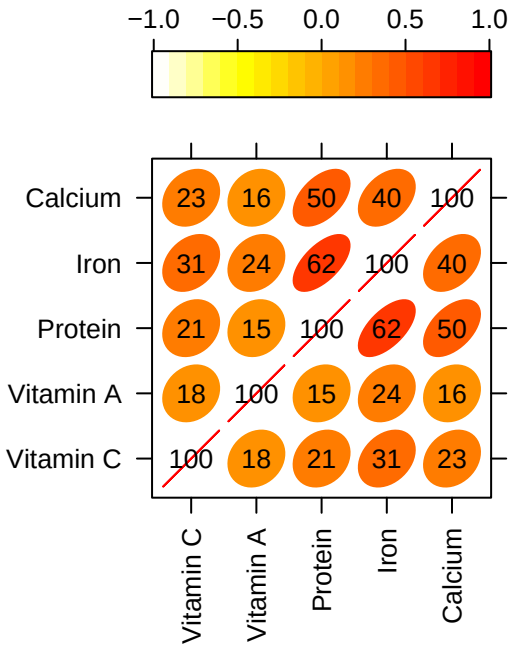
```
cor(nutrient)
```

```
##             Calcium      Iron   Protein Vitamin A Vitamin C
## Calcium   1.0000000 0.3954301 0.5001882 0.1578060 0.2292111
## Iron      0.3954301 1.0000000 0.6233662 0.2437905 0.3126009
## Protein   0.5001882 0.6233662 1.0000000 0.1467574 0.2120670
## Vitamin A 0.1578060 0.2437905 0.1467574 1.0000000 0.1835227
## Vitamin C 0.2292111 0.3126009 0.2120670 0.1835227 1.0000000
```

Looking at the pairwise plot above, there seems to be no pair of features that significantly stands out in terms of correlation. The two features with the most strong correlation seems to be iron and protein. The r correlation values from the correlation matrix confirm this, showing a r correlation = .6233, which is the largest magnitude of the feature pairs. This may give us insight as to how the PCA and MLE analysis will weigh the features in the principal components and factors.

To make it easier for non-techninal audience to understand the relationship between each variable, we also use a level plot.

```
panel.corrgram <- function(x, y, z, subscripts, at, level = 0.9, label = FALSE, ...) {
     require("ellipse", quietly = TRUE)
     x <- as.numeric(x)[subscripts]
     y <- as.numeric(y)[subscripts]
     z <- as.numeric(z)[subscripts]
     zcol <- level.colors(z, at = at,  ...)
     for (i in seq(along = z)) {
       ell=ellipse(z[i], level = level, npoints = 50,
                   scale = c(.2, .2), centre = c(x[i], y[i]))
       panel.polygon(ell, col = zcol[i], border = zcol[i], ...)}
     if (label){ panel.text(x = x, y = y, lab = 100 * round(z, 2), cex = 0.8,
                            col = ifelse(z < 0, "white", "black"))}}
cor_df = cor(nutrient) #correlation matrix of dataset
print(levelplot(cor_df[seq(5,1), seq(5,1)], at = do.breaks(c(-1.01, 1.01), 20), xlab = NULL,
                ylab = NULL, colorkey = list(space = "top"), panel = panel.corrgram, label = TRUE,
                col.regions=rev(heat.colors(100)), scales = list(x = list(rot = 90))))
```



An excellent method of displaying the correlations of all pairwise features is through a levelplot, as shown above. This method is especially helpful to those foreign to the technicalities of statistical analysis. You can observe all findings we discussed above in a graphic and concise manner. The reder the color, the more positive the relationship. The flatter the circle, the strong the relationship.

2. Fit the factor model using both PCA and MLE, and compare the parameter estimates. Discuss the underlying assumptions for each method. Which results do you prefer, and why?

PCA Analysis The following is the code for PCA analysis.

```
pca_fit = prcomp(nutrient, scale = TRUE) # PCA
pca_var = pca_fit$sdev^2 # eigenvalue
pca_pve = pca_var/sum(pca_var) # proportion explained
pca_out = cbind(pca_var, pca_pve, cumsum(pca_pve))
colnames(pca_out) = c("Eigenvalue","Proportion","Cumulative")
rownames(pca_out) = c("PC1","PC2","PC3","PC4","PC5")
pca_out
```

```
##     Eigenvalue Proportion Cumulative
## PC1  2.2812550 0.45625099  0.4562510
## PC2  0.9539042 0.19078083  0.6470318
## PC3  0.8036539 0.16073078  0.8077626
## PC4  0.6184136 0.12368272  0.9314453
## PC5  0.3427734 0.06855467  1.0000000
```

MLE Factor Analysis The following is the code for MLE factor analysis.

```
n.factors <- 2 # highest number of factors applicable to the data
mle_fit <- factanal(nutrient, n.factors, rotation="varimax") # MLE
print(mle_fit$loadings, cutoff=.05) # prop of var & loading
```

```
##
## Loadings:
##           Factor1 Factor2
## Calcium   0.466   0.298
## Iron      0.568   0.474
## Protein   0.989   0.131
## Vitamin A 0.098   0.378
## Vitamin C 0.151   0.479
##
##                Factor1 Factor2
## SS loadings       1.55   0.703
## Proportion Var    0.31   0.141
## Cumulative Var    0.31   0.451
```

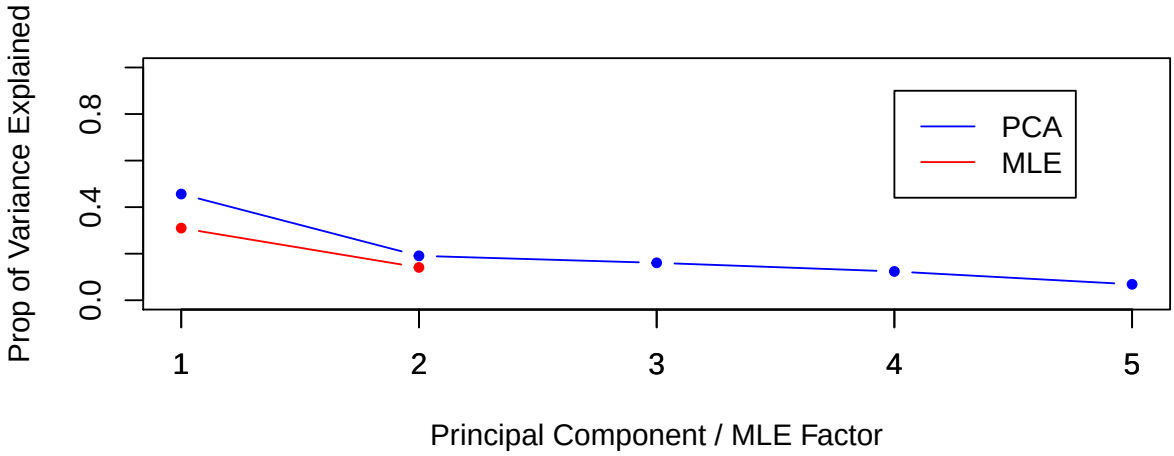
The assumption for primary component analysis is that the variances are continuous, and that there is a linear relationship between all variables. Moreover, we also need a large enough sample, of which 737 is more than enough. Finally, our data must have enough variables to reduce, and have no significant outliers. All of these are true of our dataset except for the assumption regarding outliers, thus results must account for that discrepancy.

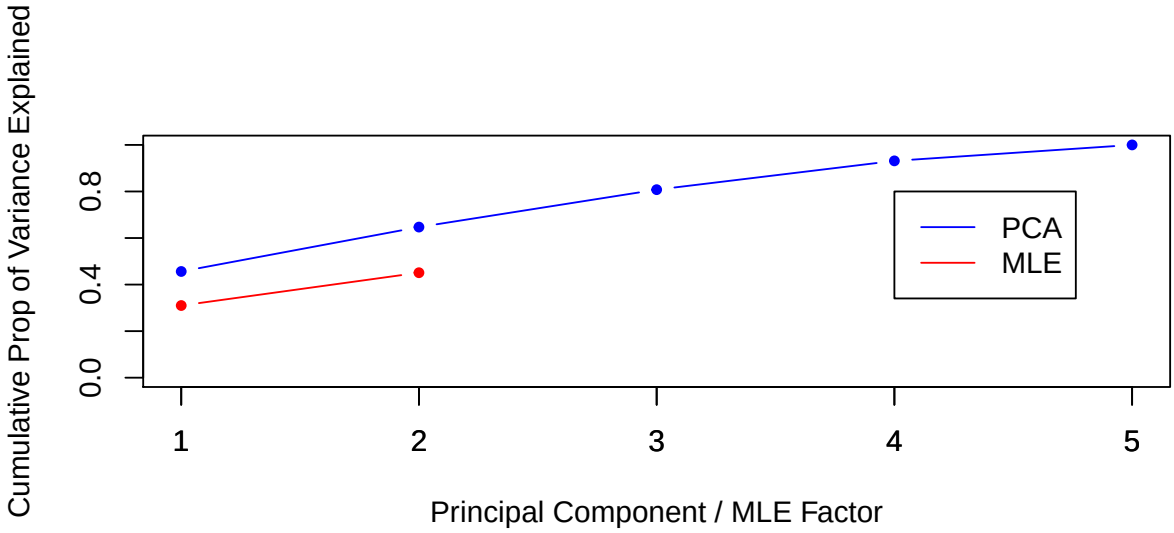
Meanwhile, for maximum likelihood estimation, we assume that the data is identically and independently distributed. However, neither of these are the case, as foods can contain multiple of these nutrients simultaneously, and thus they cannot be independent of each other, while their variances are also unequal, and thus they cannot be identically distributed.

Given the previously stated assumptions, using PCA is preferable, as only one of five underlying assumptions are violated while all MLE assumptions are violated. We will also explore the results (proportion of variance, loadings, and factor scores) of both PCA and MLE below and discuss which methods seem to perform better.

3. Use a scree plot to decide on a dimension reduction, and justify your choice.

```
par(mfrow = c(2,1)) # format the graphs
mle_pve <- c(0.31, 0.141)
# scree plots & cumulative proportion plot
plot(pca_pve, xlab="Principal Component / MLE Factor", col="blue",
     ylab="Prop of Variance Explained", ylim=c(0,1), type='b', pch=20)
lines(mle_pve, type='b', pch=20, col="red")
axis(1, at=c(1,2,3,4,5), labels=c(1,2,3,4,5))
legend(4, 0.9, legend=c("PCA", "MLE"), col=c("blue", "red"), lty = 1)
plot(cumsum(pca_pve), xlab="Principal Component / MLE Factor", col="blue",
     ylab="Cumulative Prop of Variance Explained", ylim=c(0,1), type='b', pch=20)
lines(cumsum(mle_pve), type='b', pch=20, col="red")
axis(1, at=c(1,2,3,4,5), labels=c(1,2,3,4,5))
legend(4, 0.8, legend=c("PCA", "MLE"), col=c("blue", "red"), lty = 1)
```





Looking at the variance explained by each principal component, we can see that even the first component explains only around 45.62% followed by 19.07% variance explained from the second principal component. The amount of variance diminishes after the second component. The scree plot above displays this pattern. As a result, we believe that only two components should be utilized.

Also, for MLE factor analysis, the factor model could not be fit for more than 2 factors, with the first factor explaining 31% of the variance and the second factor explaining 14.1% of the variance.

Compared with MLE, the first two components generated by PCA (45.62%, 19.07%) explained more variance than the first two factors generated by MLE (31%, 14.1%). Thus, we decide to use PCA as our method for dimension reduction.

4. Examine the factor loadings, and discuss in your report which variables have high or low loadings. Can you associate an interpretation to your factors?

```
t(pca_fit$rotation[,1:2]) # PCA loading
```

```
##        Calcium        Iron    Protein Vitamin A Vitamin C
## PC1  0.4725630  0.54314812  0.5370491 0.2724785 0.3449756
## PC2 -0.2658644 -0.09248598 -0.3476750 0.7825987 0.4329248
```

We decide to only interpret the first two PCs from PCA results since the first two PCs already explained over 60% of the total variance. From the loading results above from PCA, we can see that calcium, iron, and protein have relatively higher loadings on the first principal component, while vitamin A and vitamin C have higher/positive loadings on the second principal component compared with the negative loadings of calcium, iron, and protein on the second principal component.

From the loadings of the first two components, it seems that the five variables are separated into two latent groups. One group contains calcium, iron, and protein, while the other contains vitamin A and vitamin C. More of the potential reasons will be discussed below.

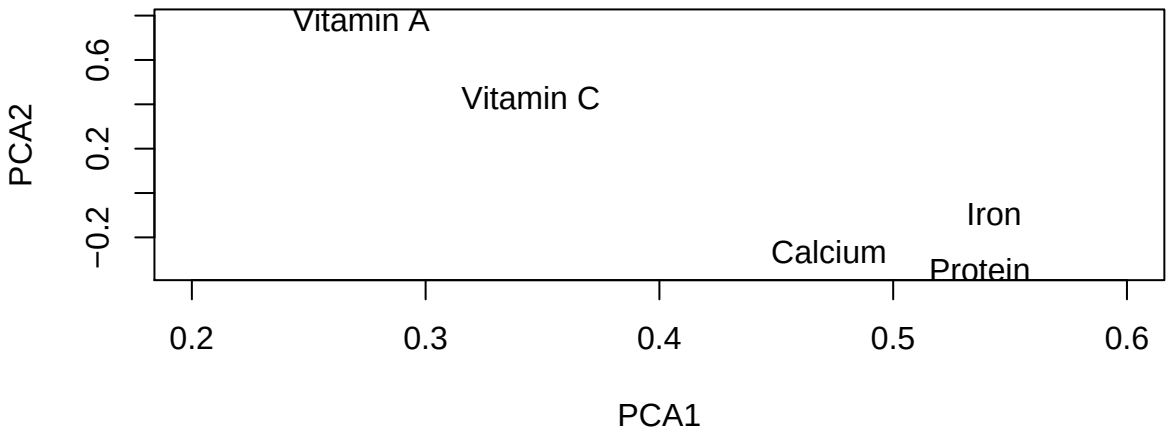
```
t(mle_fit$loadings[,1:2]) # MLE loading
```

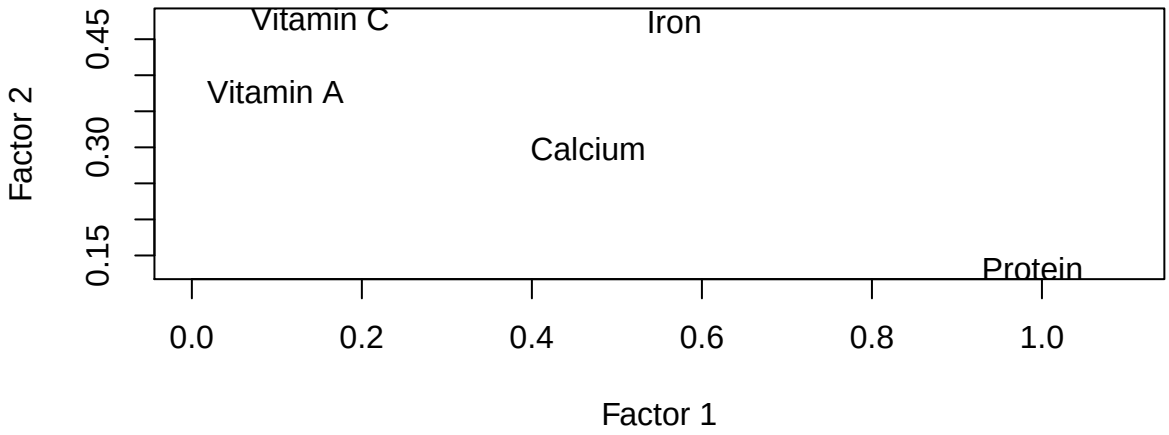
```
##           Calcium      Iron   Protein  Vitamin A Vitamin C
## Factor1 0.4662298 0.5675046 0.9888518 0.09839555 0.1510572
## Factor2 0.2984038 0.4743206 0.1310440 0.37773096 0.4787664
```

A similar finding can actually be found also in the first factor of the MLE factor analysis, with protein (highest), iron, and calcium having high loadings, while vitamin A and vitamin C have low loadings on the first factor. The second factor from MLE seems less interpretable since now iron, vitamin A and vitamin C has higher loadings, while calcium and protein has lower loadings. This is also part of the reason why we want to choose PCA over factor analysis.

5. Examine the factor scores by scatter plots or pairwise scatter plots. Is there a story to tell from these results?

```
par(mfrow = c(2,1)) # format the graphs
# scatter plot of PCA1 & PCA2
pca_loading = pca_fit$rotation[,1:2]
plot(pca_loading,type="n", xlim = c(.2,.6),xlab = "PCA1", ylab = "PCA2")
text(pca_loading,labels=names(nutrient))
# scatterplot of Factor 1 & Factor 2
mle_loading = mle_fit$loadings[,1:2]
plot(mle_loading,type="n", xlim = c(0,1.1),xlab = "Factor 1", ylab = "Factor 2")
text(mle_loading,labels=names(nutrient))
```





In the MLE factor analysis, we can see that iron and calcium seem to for a latent group, vitamin A and vitamin C seem to for another, and protein seems to be on its own (this is not our primary focus). In the PCA results, by factor scores, we can see that calcium, iron, and protein are better primary components to analyze with than vitamin A and vitamin C. There is a lot less variance explained by the vitamins, and from there we can conclude that the people represented by this dataset consume a relatively more consistent amount of those vitamins, while the amount of calcium, iron, and protein digested varies much more. This may be explained by the foods that people eat, as discussed in the conclusion.

Conclusion & Discussion

Looking at our results, we see that there are two latent groups between the nutritional features. One group includes: Calcium, Iron, and Protein. While the other group includes: Vitamin A and Vitamin C. One

distinguishing feature between these two groups is that calcium, iron and protein are classified as minerals, which are inorganic elements. Vitamin C and vitamin A are classified as organic substances.

The classification of organic and inorganic nutrition features may explain the latent grouping, but it is also possible that certain commonly eaten foods include higher concentrations of calcium, iron and protein, over other organic vitamins. It is also known that highly processed foods may contain less organic vitamins, and with the growing popularity of processed foods, we could expect a higher presence of inorganic nutrients.

Alternatively, without the processed food lens, it is possible that the unequal amounts of nutrients found naturally within foods can explain the PCA. Calcium is particularly rich in dairy, and while not the only source of calcium, conditions like lactose intolerance leads to high variance in the amount of dairy (and thus calcium) that an average person may consume. Meanwhile, vitamin A is found much more evenly across all vegetables and meats, and that could be a reason why that component is less relevant in analysis.

Appendix

Principal Component Analysis (PCA) [adapted from lecture 19 & 20]

What is PCA?

- PCA is a widely-used approach for extracting a small set of features from a large set of variables, meanwhile retaining most of the information.
- PCA is a mathematical procedure that transforms many possibly correlated variables into a smaller number of uncorrelated variables called *principal components*.
- The first principal component accounts for as much of the variability in the data as possible, and each succeeding principal component accounts for as much of the remaining variability as possible.

How to find PCs?

- The principal directions can be found by computing the eigenvalues and eigenvectors of the covariance matrix $\mathbf{\Sigma}$.
- Let $\lambda_1, \ldots, \lambda_p$ denote the eigenvalues of $\mathbf{\Sigma}$ in descending order, i.e.

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p.$$

- Let $\underline{e}_1, \ldots, \underline{e}_p$ be the corresponding eigenvectors (with unit length):

$$\mathbf{\Sigma}\, \underline{e}_j = \lambda_j\, \underline{e}_j$$

  for $j = 1, \ldots, p$.
- In fact, the jth eigenvector will serve as the jth principal direction!

1. $\underline{e}_j' \underline{e}_j = 1$ and $\underline{e}_j' \underline{e}_k = 0$ for $k \neq j$, due to orthogonality of eigenvectors (if eigenvalues are distinct).
2. Because
$$\mathrm{Var}[y_j] = \underline{e}_j' \mathbf{\Sigma} \underline{e}_j = \lambda_j \underline{e}_j' \underline{e}_j = \lambda_j,$$
we see that $\mathrm{Var}[y_1] \geq \mathrm{Var}[y_2] \geq \ldots \geq \mathrm{Var}[y_p]$.
3. Also for $j \neq k$
$$\mathrm{Cov}[y_j, y_k] = \underline{e}_j' \mathbf{\Sigma} \underline{e}_k = \lambda_k \underline{e}_j' \underline{e}_k = 0$$
so the new variables $y_1, \ldots, y_p$ are uncorrelated.

Maximum Likelihood Estimation (MLE) [adapted from lecture 21 & 22]

What is MLE Factor Analysis?

- Factor Analysis is a method for modeling observed variables, and their covariance structure, in terms of a smaller number of underlying unobservable (latent) **factors**.
- Factor analysis is a very flexible dimension reduction method. It is similar to PCA, but more elaborate. In some sense, factor analysis can be considered as an "inversion" of PCA.
- In factor analysis, we model the observed variables as linear functions of the **factors**.

How to find factors?

- Under the iid normal assumption, the joint density of $\underline{x}_1, \ldots, \underline{x}_n$ is

$$f(\underline{x}_1, \ldots, \underline{x}_n) = \prod_{i=1}^{n} f(\underline{x}_i) = (2\pi)^{-np/2} \det \mathbf{\Sigma}^{-n/2} \exp\{-\frac{1}{2} \sum_{i=1}^{n} (\underline{x}_i - \underline{\mu})' \mathbf{\Sigma}^{-1} (\underline{x}_i - \underline{\mu})\}.$$

- Then, the log-likelihood function of $\underline{\mu}$, $\mathbf{L}$, and $\mathbf{\Psi}$ is (up to a constant, which we can ignore)

$$\ell_n(\underline{\mu}, \mathbf{L}, \mathbf{\Psi}) = -\frac{n}{2} \log \det[\mathbf{L}\,\mathbf{L}' + \mathbf{\Psi}] - \frac{1}{2} \sum_{i=1}^{n} (\underline{x}_i - \underline{\mu})' (\mathbf{L}\,\mathbf{L}' + \mathbf{\Psi})^{-1} (\underline{x}_i - \underline{\mu}).$$

- The maximum likelihood estimator for the mean vector $\underline{\mu}$, the factor loadings $\mathbf{L}$, and the specific variances $\mathbf{\Psi}$ are obtained by finding $\hat{\underline{\mu}}$, $\hat{\mathbf{L}}$, and $\hat{\mathbf{\Psi}}$ that maximize the log-likelihood $\ell_n(\underline{\mu}, \mathbf{L}, \mathbf{\Psi})$.
- In general, there is no closed-form solution to this maximization problem. In practice, it is usually solved by iterative optimization algorithms.

Works & Codes Cited

1. McElroy, Tucker. https://github.com/tuckermcelroy/ma189/blob/main/Lectures/Ma189Lecture19.Rmd
2. McElroy, Tucker. https://github.com/tuckermcelroy/ma189/blob/main/Lectures/Ma189Lecture20.Rmd
3. McElroy, Tucker. https://github.com/tuckermcelroy/ma189/blob/main/Lectures/Ma189Lecture21.Rmd
4. McElroy, Tucker. https://github.com/tuckermcelroy/ma189/blob/main/Lectures/Ma189Lecture22.Rmd