FINM 331: MULTIVARIATE DATA ANALYSIS
FALL 2021
PROBLEM SET 3

The required files for all problems can be found either in the in the subfolder `hw3` under 'Files' in Canvas or at the following URL:

`http://www.stat.uchicago.edu/~lekheng/courses/331/hw3/`

You are welcomed to use any programming language or software packages you like.

1. (*Factor Analysis*) This is the same air quality data set we saw in the previous problem set but this time we will only take four variables X_1, X_2, X_5 and X_6 by leaving out `CO`, `NO`, and `HC` variables.

(a) Obtain the principal component solution to the factor model $\mathbf{X} = \boldsymbol{\mu} + L\mathbf{F} + \boldsymbol{\varepsilon}$ with number of factors $m = 1$ and $m = 2$ using:
  (i) the sample covariance matrix;
  (ii) the sample correlation matrix.

In other words, you should find the matrix factor loadings $L \in \mathbb{R}^{n\times m}$, the specific variances $\psi_1, \dots, \psi_p \in \mathbb{R}$, and write down the proportions of variability (in percentages) due to the factors.

(b) Find the angle between the first factor loading in (i) and that the first factor loading in (ii).

(c) For the $m = 2$ case, compare the factor loadings obtained in (i) and that in (ii) using orthogonal Procrustes analysis.

(d) Comment on your results.

2. (*Population Canonical Correlation Analysis*) The 2×1 random vectors $\mathbf{X}$ and $\mathbf{Y}$ have joint mean vector and joint covariance matrix

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \end{bmatrix} \in \mathbb{R}^{4\times 4}, \quad \Sigma = \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix} \in \mathbb{R}^{4\times 4},$$

where

$$\boldsymbol{\mu}_X = \begin{bmatrix} -3 \\ 2 \end{bmatrix}, \quad \boldsymbol{\mu}_Y = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

and

$$\Sigma_X = \begin{bmatrix} 8 & 2 \\ 2 & 5 \end{bmatrix}, \quad \Sigma_Y = \begin{bmatrix} 6 & -2 \\ -2 & 7 \end{bmatrix}, \quad \Sigma_{YX}^{\mathsf{T}} = \Sigma_{XY} = \begin{bmatrix} 3 & 1 \\ -1 & 3 \end{bmatrix}.$$

(a) Calculate the canonical correlation ρ_1 (the largest), ρ_2 (the second largest).

(b) Find the canonical correlation variables (U_1, V_1) and (U_2, V_2) corresponding to ρ_1 and ρ_2.

(c) Let $U = [U_1, U_2]^{\mathsf{T}}$ and $V = [V_1, V_2]^{\mathsf{T}}$. Evaluate

$$E\left(\begin{bmatrix} U \\ V \end{bmatrix}\right) \quad \text{and} \quad \operatorname{Cov}\left(\begin{bmatrix} U \\ V \end{bmatrix}\right) = \begin{bmatrix} \Sigma_U & \Sigma_{UV} \\ \Sigma_{VU} & \Sigma_V \end{bmatrix}$$

(d) Comment on the correlation structure between and within U and V.

3. (*Sample canonical correlation analysis*) The data set for this problem is obtained by taking four different measures of stiffness, `shock`, `vibrate`, `static1`, `static2`, for each of $n = 30$ boards. The first measurement involves sending a shock wave down the board, the second measurement

Date: November 8, 2021 (Version 1.0); due: November 21, 2021.

is determined while vibrating the board, and the last two measurements are obtained from static tests. The squared distances $d_j^2 = (\mathbf{x}_j - \overline{\mathbf{x}})^\intercal S^{-1}(\mathbf{x}_j - \overline{\mathbf{x}})$ are also included as the last column in the data matrix.

Let $\mathbf{X} = [X_1, X_2]^\intercal$ be the random vector representing the dynamic measures of stiffness, and let $\mathbf{Y} = [Y_1, Y_2]^\intercal$ be the random vector representing the static measures of stiffness. Load the data matrix `p3.txt` (R command: `stiff = read.table("p3.txt")`)

(a) Perform a canonical correlation analysis of these data by computing the singular value decomposition of an appropriate matrix formed from the sample covariance matrices. You may compare your result with that obtained by your software (if you use R, it is `cancor(X1,X2)`).

(b) Write the first canonical correlation variables U_1 and V_1 as linear combinations of `shock`, `vibrate`, `static1`, `static2`.

(c) Produce two scatterplots of the data: one in the coordinate plane of the first canonical correlation vectors, one in the plane of the second canonical correlation vectors.

(d) Based on the two plots and the values of the canonical correlations $\{\rho_1, \rho_2\}$, comment on the correlation structure captured by each canonical pair.

(e) Repeat (a) with sample correlation matrices in place of sample covariance matrices and verify that the pairs of canonical vectors obtained are related via scaling by the sample standard deviation matrix.

4. (*Canonical correlation analysis for angular measurements*) Some observations are in the form of angles. Here we will see how to deal with such data.

(a) Consider a bivariate random vector $\mathbf{X} = [X_1, X_2]^\intercal$ with a uniform distribution inside a circle of radius 1 centered at some unknown point

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \in \mathbb{R}^2.$$

Then $E(\mathbf{X}) = \boldsymbol{\mu}$. A sample of $n = 4$ is taken. The observed values are

$$\begin{bmatrix} 0.9 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0.6 \\ 0.6 \end{bmatrix}, \quad \begin{bmatrix} 0.6 \\ -0.6 \end{bmatrix}, \quad \begin{bmatrix} -0.9 \\ 0 \end{bmatrix}.$$

Compute sample mean $\overline{\mathbf{x}}$ and sample covariance matrix. Is $\overline{\mathbf{x}}$ a good estimator of $\boldsymbol{\mu}$? Why or why not?

(b) We consider an angular valued random variable θ, note that this can always be represented as a random vector $\mathbf{Y} = [\cos\theta, \sin\theta]^\intercal$ that takes value on the circle. Show that

$$\mathbf{b}^\intercal \mathbf{Y} = \sqrt{b_1^2 + b_2^2}\cos(\theta - \beta)$$

where $b_1/\sqrt{b_1^2 + b_2^2} = \cos\beta$ and $b_2/\sqrt{b_1^2 + b_2^2} = \sin\beta$. Here $\mathbf{b} = [b_1, b_2]^\intercal \in \mathbb{R}^2$ is a constant vector.

(c) Let $\mathbf{X} = X$ be a random vector with a single component, i.e., just a random variable. Here X is not angular valued. Show that the population canonical correlation is

$$\rho_1 = \max_\beta \operatorname{Corr}\big(X, \cos(\theta - \beta)\big)$$

and that selecting the population canonical correlation variable V_1 amounts to selecting a new 'origin' or 'baseline' β for the angle θ.

(d) Let X is a random variable representing ozone (O_3) levels and θ is a angular random variable representing wind direction measured from the north. We make 19 observations to obtain the sample correlation matrix

$$R = \begin{bmatrix} R_X & R_{X\theta} \\ R_{\theta X} & R_\theta \end{bmatrix} = \begin{array}{c|ccc} & \mathrm{O}_3 & \cos\theta & \sin\theta \\ \hline \mathrm{O}_3 & 1.000 & 0.166 & 0.694 \\ \cos\theta & 0.166 & 1.000 & -0.051 \\ \sin\theta & 0.694 & -0.051 & 1.000 \end{array}.$$

Find the sample canonical correlation $\widehat{\rho}_1$ and the sample canonical correlation variable $\widehat{V}_1$ representing the new origin β.

(e) Let ϕ be another angular valued random variable and let $\mathbf{X} = [\cos\phi, \sin\phi]^\intercal$. Then similar to (b), we get

$$\mathbf{a}^\intercal \mathbf{X} = \sqrt{a_1^2 + a_2^2}\cos(\phi - \alpha).$$

Now show that

$$\rho_1 = \max_{\alpha,\beta} \operatorname{Corr}\big(\cos(\phi-\alpha), \cos(\theta-\beta)\big).$$

(f) Let ϕ and θ be two angular random variables representing wind directions at 6:00 A.M. and at 12:00 P.M. We make 21 measurements of $\mathbf{X}$ and $\mathbf{Y}$ (related to ϕ and θ as in (b) and (d)). We obtain the sample correlation matrix

$$R = \begin{bmatrix} R_X & R_{XY} \\ R_{YX} & R_Y \end{bmatrix} = \begin{array}{c|cccc} & \cos\phi & \sin\phi & \cos\theta & \sin\theta \\ \hline \cos\phi & 1.000 & -0.291 & 0.440 & 0.372 \\ \sin\phi & -0.291 & 1.000 & -0.205 & 0.243 \\ \cos\theta & 0.440 & -0.205 & 1.000 & 0.181 \\ \sin\theta & 0.372 & 0.243 & 0.181 & 1.000 \end{array}.$$

Find the sample canonical correlation $\widehat{\rho}_1$ and sample canonical correlation variables $\widehat{U}_1$ and $\widehat{V}_1$.

5. (*Proofs behind* CCA) Let $A \in \mathbb{R}^{p\times p}$ and $B \in \mathbb{R}^{q\times q}$ be symmetric positive definite matrices and $C \in \mathbb{R}^{p\times q}$. Let

$$G = A^{-1/2}CB^{-1/2} \in \mathbb{R}^{p\times q}.$$

We shall write $\lambda_{\max}(M)$ for the largest eigenvalue of a matrix M.

(a) Suppose $p = q$. Show that eigenvalues of $B^{-1}A$, $B^{-1/2}AB^{-1/2}$, and AB^{-1} are all equal. What are the relations between the eigenvectors?

(b) Suppose $p = q$. Show that

$$\max_{\mathbf{x}\in\mathbb{R}^p}\{\mathbf{x}^\intercal A\mathbf{x} : \mathbf{x}^\intercal B\mathbf{x} = 1\} = \max_{\mathbf{y}\in\mathbb{R}^p}\{\mathbf{y}^\intercal B^{-1/2}AB^{-1/2}\mathbf{y} : \mathbf{y}^\intercal\mathbf{y} = 1\}.$$

By using Problem **8** in Homework **2**, deduce that

$$\max_{\mathbf{x}\in\mathbb{R}^p}\{\mathbf{x}^\intercal A\mathbf{x} : \mathbf{x}^\intercal B\mathbf{x} = 1\} = \lambda_{\max}(B^{-1/2}AB^{-1/2}),$$
$$\operatorname*{argmax}_{\mathbf{x}\in\mathbb{R}^p}\{\mathbf{x}^\intercal A\mathbf{x} : \mathbf{x}^\intercal B\mathbf{x} = 1\} = \mathbf{q}_{\max},$$

where $\mathbf{q}_{\max} \in \mathbb{R}^p$ is the eigenvector of $B^{-1}A$ corresponding to the largest eigenvalue.

(c) Show that if we fix $\mathbf{x} \in \mathbb{R}^p$ and just maximize over all $\mathbf{y} \in \mathbb{R}^q$, then

$$\max_{\mathbf{y}\in\mathbb{R}^q}\{(\mathbf{x}^\intercal C\mathbf{y})^2 : \mathbf{y}^\intercal B\mathbf{y} = 1\} = \max_{\mathbf{y}\in\mathbb{R}^q}\{\mathbf{y}^\intercal[C^\intercal\mathbf{x}\mathbf{x}^\intercal C]\mathbf{y} : \mathbf{y}^\intercal B\mathbf{y} = 1\}$$

and deduce that from (a) and (b) that

$$\max_{\mathbf{y}\in\mathbb{R}^q}\{(\mathbf{x}^\intercal C\mathbf{y})^2 : \mathbf{y}^\intercal B\mathbf{y} = 1\} = \lambda_{\max}(B^{-1}C^\intercal\mathbf{x}\mathbf{x}^\intercal C).$$

Show that the largest eigenvalue of a rank-1 matrix $\mathbf{a}\mathbf{b}^\intercal$ is $\mathbf{b}^\intercal\mathbf{a}$ and deduce that

$$\max_{\mathbf{y}\in\mathbb{R}^q}\{(\mathbf{x}^\intercal C\mathbf{y})^2 : \mathbf{y}^\intercal B\mathbf{y} = 1\} = \mathbf{x}^\intercal C B^{-1} C^\intercal \mathbf{x}.$$

(d) Using (a), (c), and Problem **7** in Homework **2**, show that

$$\max_{\mathbf{x}\in\mathbb{R}^p,\ \mathbf{y}\in\mathbb{R}^q}\{(\mathbf{x}^\intercal C\mathbf{y})^2 : \mathbf{x}^\intercal A\mathbf{x} = 1,\ \mathbf{y}^\intercal B\mathbf{y} = 1\} = \lambda_{\max}(GG^\intercal).$$

(e) Let $\sigma_1,\dots,\sigma_p \in \mathbb{R}$, $\mathbf{u}_1,\dots,\mathbf{u}_p \in \mathbb{R}^p$, $\mathbf{v}_1,\dots,\mathbf{v}_p \in \mathbb{R}^q$ be the singular values and left/right singular vectors of G. By Problem **7** in Homework **2**, show that

$$\max_{\mathbf{x}\in\mathbb{R}^p}\{\mathbf{x}^\intercal GG^\intercal\mathbf{x} : \mathbf{x}^\intercal\mathbf{x} = 1,\ \mathbf{u}_i^\intercal\mathbf{x} = 0,\ i = 1,\dots,k-1\} = \sigma_k^2,$$
$$\operatorname*{argmax}_{\mathbf{x}\in\mathbb{R}^p}\{\mathbf{x}^\intercal GG^\intercal\mathbf{x} : \mathbf{x}^\intercal\mathbf{x} = 1,\ \mathbf{u}_i^\intercal\mathbf{x} = 0,\ i = 1,\dots,k-1\} = \mathbf{u}_k,$$

for $k = 1,\dots,p$. Hence deduce that

$$\max_{\mathbf{x}\in\mathbb{R}^p,\ \mathbf{y}\in\mathbb{R}^q}\{\mathbf{x}^\intercal C\mathbf{y} : \mathbf{x}^\intercal A\mathbf{x} = 1,\ \mathbf{y}^\intercal B\mathbf{y} = 1,\ \mathbf{u}_i^\intercal A^{1/2}\mathbf{x} = 0,\ i = 1,\dots,k-1\} = \sigma_k,$$
$$\operatorname*{argmax}_{\mathbf{x}\in\mathbb{R}^p,\ \mathbf{y}\in\mathbb{R}^q}\{\mathbf{x}^\intercal C\mathbf{y} : \mathbf{x}^\intercal A\mathbf{x} = 1,\ \mathbf{y}^\intercal B\mathbf{y} = 1,\ \mathbf{u}_i^\intercal A^{1/2}\mathbf{x} = 0,\ i = 1,\dots,k-1\} = (A^{-1/2}\mathbf{u}_k, B^{-1/2}\mathbf{v}_k),$$

for $k = 1,\dots,p$. Finally show that

$$\max_{\mathbf{x}\in\mathbb{R}^p,\ \mathbf{y}\in\mathbb{R}^q}\{\mathbf{x}^\intercal C\mathbf{y} : \mathbf{x}^\intercal A\mathbf{x} = 1,\ \mathbf{y}^\intercal B\mathbf{y} = 1,\ \mathbf{u}_i^\intercal A^{1/2}\mathbf{x} = 0,\ \mathbf{v}_i^\intercal B^{1/2}\mathbf{y} = 0,\ i = 1,\dots,k-1\} = \sigma_k,$$
$$\operatorname*{argmax}_{\mathbf{x}\in\mathbb{R}^p,\ \mathbf{y}\in\mathbb{R}^q}\{\mathbf{x}^\intercal C\mathbf{y} : \mathbf{x}^\intercal A\mathbf{x} = 1,\ \mathbf{y}^\intercal B\mathbf{y} = 1,\ \mathbf{u}_i^\intercal A^{1/2}\mathbf{x} = 0,\ \mathbf{v}_i^\intercal B^{1/2}\mathbf{y} = 0,\ i = 1,\dots,k-1\} = (A^{-1/2}\mathbf{u}_k, B^{-1/2}\mathbf{v}_k),$$

for $k = 1,\dots,p$.

6. (*Linear discriminant analysis*) The admissions committee of a business school used GPA and GMAT scores to make admission decisions. The values for the variable `admit = 1,2,3` correspond to admission decisions of `yes`, `no`, `waitlist`. Label the data set `p6.txt` — helpful R commands:

```
gsbdata = read.table("p6.txt"); colnames(gsbdata)=c("GPA", "GMAT","admit");
```

(a) Calculate $\overline{\mathbf{x}}_i, \overline{\mathbf{x}}$ and S_{pool}.

(b) Calculate the sample within groups matrix W, its inverse W^{-1}, and the sample between groups matrix B. Find the eigenvalues and eigenvectors of $W^{-1}B$. (R command for A^{-1} is `solve(A)`).

(c) Use the linear discriminants derived from these eigenvectors to classify the two new observations $\mathbf{x} = [3.21, 497]^\intercal$ and $\mathbf{x} = [3.22, 497]^\intercal$.

(d) Scatterplot the original data set on the plane of the first two discriminants, labeled by admission decisions. Comment on the results in (c). Is this a good admission policy?

7. (*Correspondence Analysis*) A client of a law firm would like to visualize the number of large class-action lawsuits each year across different industries from 2011 to the first half of 2017. The correspondence analysis provides a means of displaying or summarizing a set of categorical data in two-dimensional graphical form. The data on class-action lawsuits are from annual reports of Stanford Law School's Securities Class Action Clearinghouse. To load the data in R, you can use the following command:

```
CALaw = read.csv("/classaction_lawsuit.csv", header=TRUE)
```

Notation: Denote X as a data matrix of the number of class action lawsuits for industry-year. $x_{i\bullet}$ denotes row total (summing across all years for each industry). $x_{\bullet j}$ denotes column total (summing across all industries for each year). $x_{\bullet\bullet}$ denotes grand total. Define $D_r = \text{diag}(x_{1\bullet},\dots,x_{n\bullet})$ and $D_c = \text{diag}(x_{\bullet 1},\dots,x_{\bullet p})$.

(a) What are the dimensions, n and p, in this dataset?

(b) Show 1 is an eigenvalue of matrices $D_r^{-1}XD_c^{-1}X^\intercal$ and $D_c^{-1}X^\intercal D_r^{-1}X$ and that the corresponding eigenvectors are proportional to $\mathbf{1} = [1, \dots, 1]^\intercal$.

(c) Transform the data as follows:

$$Y = \sqrt{x_{\bullet\bullet}}D_r^{-1/2}\Big(X - \frac{ab^\intercal}{x_{\bullet\bullet}}\Big)D_c^{-1/2} \in \mathbb{R}^{n\times p},$$

where $a = D_r\mathbf{1}_n$ and $b = D_c\mathbf{1}_p$. Report the SVD on Y (both singular values and left/right singular vectors). Is there another formula to compute the entries of the matrix Y?

(d) Write down the formula to compute row weight vectors and column weight vectors. How many different row weight vectors and column weight vectors are there? Report all row weight vectors and column weight vectors.

(e) Emulating what we did in PCA, make the following two plots:
   (i) Scatterplot of the first two row weight vectors: Does this scatterplot inform us about year or industry? What do you learn from this scatterplot?
   (ii) Biplot: What do you learn from the biplot?

(f) Write down the formula to calculate the Frobenius norm of Y. Compute the Frobenius norm of Y. What is the relationship between the sum of squares of the singular values and the Frobenius norm of Y?

(g) Report the percentage of original variance that each dimension in the row/column weight vectors explain? How many singular values are needed to effectively summarize at least 90% of the variability in the data?

8. (*Multidimensional Scaling*) An investor looking to allocate his funds to different industries seeks to visually understand the relationship between returns across different US industries. This investor has a deep pocket but does not know statistics so he comes to seek your advice. As a financial mathematician, multidimensional scaling first comes to your mind to answer this investor's question. To collect the data, the US industry returns can be downloaded from the 'Industry Return' sections at:

`http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html#Research`

For the purpose of this problem set, the dataset of monthly returns of 30 US industries is downloaded and formatted. To read in the dataset in R, you may use

```
FF=read.csv("./FamaFrench30.csv", header=TRUE)
```

Each row of the data represents how the industry in each column goes up or down on different date. The value of 1 means that the industry on a particular column goes up 1% on that month, compared to the previous month.

(a) Report mean returns and standard deviation of five industries of your choice. Out of all 30 industries, which industry performs the best on average, which industry is the most volatile?

(b) Let R_t^i be the return of industry i at time t. Write a formula to compute the distance between two industries. Denote what each subscript/superscript means and specify dimension of each subscript/superscript (i.e., explicitly stating what do you sum to). Write a code to compute distance and report the distance of the following pair of industries:
   (i) `Autos` and `ElecEq`;
   (ii) `Autos` and `Trans`;
   (iii) `Autos` and `Oil`.

(c) Do you need to demean the data to compute distance matrix? Why?

(d) Report the distance matrix of all industries. To conveniently compute distance, R has a built in distance matrix command `dist`.

```
dist(data matrix, method = "euclidean", diag = FALSE, upper = FALSE, p = 2)
```

The first input is the data matrix. The distance command will compute the Euclidean distances among each row of the data. If you use R, you may need to convert the results into matrix using the `as.matrix` command.

(e) With the distance matrix in hand, you are now ready to perform multidimensional scaling to visualize this data. The end goal is to plot the first two dimensions after multidimensional scaling. To perform MDS, you first need Euclidean distance matrix (EDM) from the previous part. Then, you would perform the following steps
   (i) Form Gram matrix G from EDM.
   (ii) Perform EVD on G and recover X using $X = Q_p \Lambda_p^{1/2}$.

   Report the result by plotting the first two dimensions after multidimensional scaling with corresponding industry label for each data point. Does this plot have to be unique? Why?

(f) Interpret the results. What does "closer/further" in distance mean in this setting? Which industry tends to co-move with the games industry (`Games`) the most? List three industries whose returns tend to move on its own.

(g) What is your advice for an investor who put most of his money on stocks in the telecom inudstry (`Telcm`)? [*Hint*: Think of diversification]