인대현 (20223111154)

2022-03-17

HW2 Report

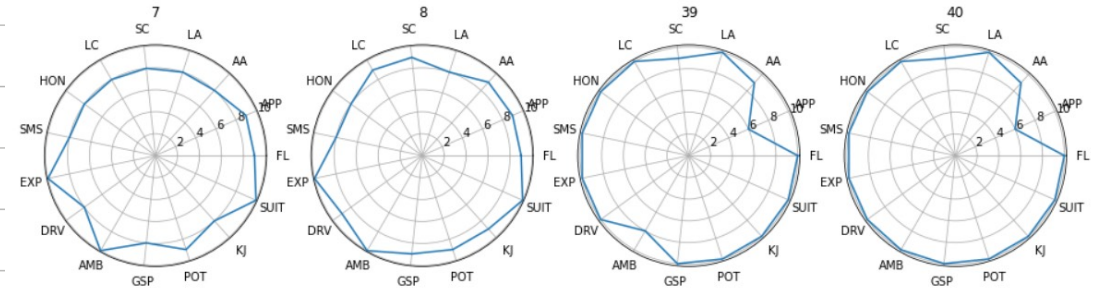
1. We plot star plots of all 48 applicants. Then, by obtaining the sum of the scores of each applicant and sorting by the sum, we can find the applicants with the highest scores as shown below:

```
In [4]: applicant2=applicant.copy()
        applicant2['Sum_Result'] = applicant2.drop(['ID'],axis=1).sum(axis = 1)
        applicant2.sort_values(by='Sum_Result',ascending=False) #In order, applicants with ID no. 40, 39, 8, and 7 have the highest score
```

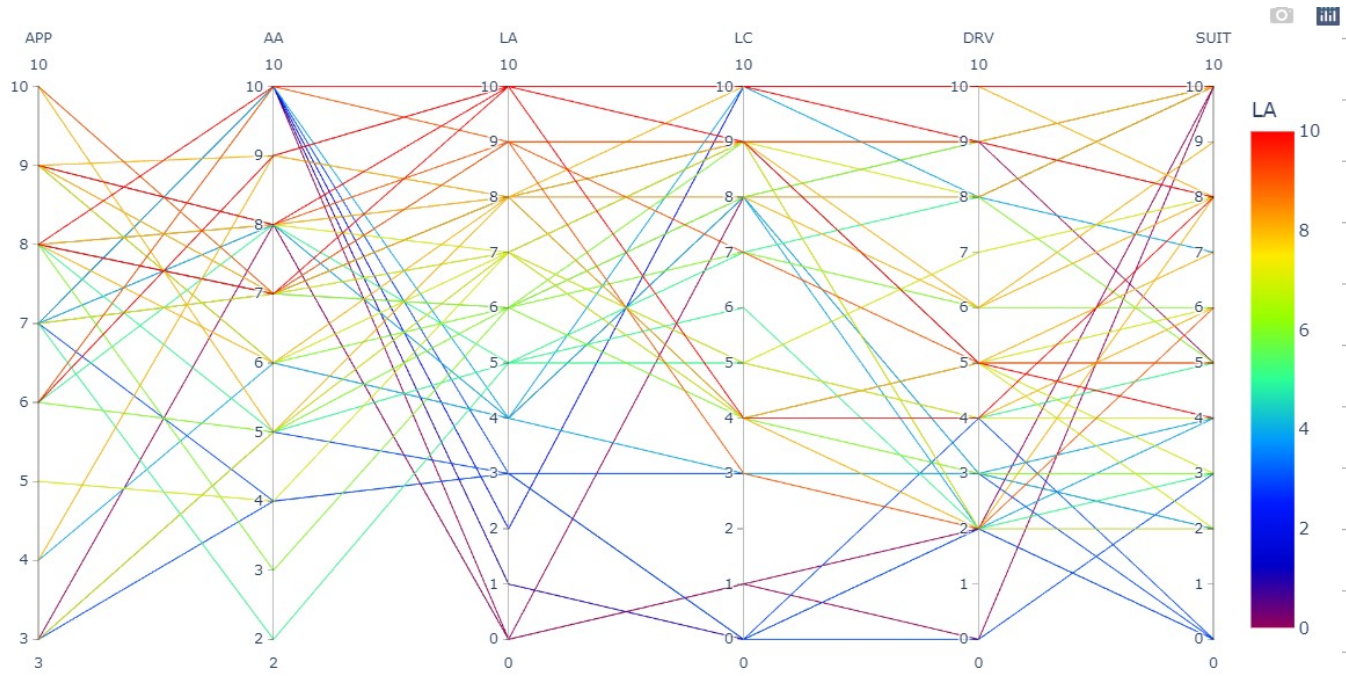
Out[4]:

	ID	FL	APP	AA	LA	SC	LC	HON	SMS	EXP	DRV	AMB	GSP	POT	KJ	SUIT	Sum_Result
39	40	10	6	9	10	9	10	10	10	10	10	10	10	10	10	10	144
38	39	10	6	9	10	9	10	10	10	10	10	8	10	10	10	10	142
7	8	9	9	9	8	9	9	8	8	10	9	10	9	9	9	10	135
6	7	9	9	8	8	8	8	8	8	10	8	10	8	9	8	10	129
22	23	7	10	7	9	9	9	10	10	3	9	9	10	9	10	8	129
21	22	9	8	7	8	9	10	10	10	3	10	8	10	8	10	8	128
1	2	9	10	5	8	10	9	9	10	5	9	9	8	8	8	10	127

Observe that applicants with ID numbers 40, 39, 8, and 7 have the top four highest scores. The star plots of the top four applicants are shown below:



2. We can create a new dataframe ('applicant3') with reindexed columns for the selected variables. We plot the parallel coordinate plot using 'applicant3' as shown below:



Applicants with lower scores of likeability (LA) tend to have high academic ability (AA). We chose variable 'LA' to highlight this inspection. It is difficult to find obvious clusters from the given data. Clusters may be found if more data is acquired.

3.(a) $\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} \underset{\sim}{x}_i\right)$ $\qquad S = \frac{1}{n-1}\left[\sum_{i=1}^{n}\left(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}}\right)\left(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}}\right)'\right]$

Using these equations we implemented the code below in order to calculate the sample mean vector and the covariance matrix of x.

In [9]:
```
#sample mean vector
mean=ones.transpose().dot(data)/len(data)
mean
```

Out[9]:

	X1	X2	X3
ones	5.0	3.0	4.0

In [10]:
```
mean_rep = pd.concat([mean]*3)
mean_rep.columns=['X1','X2','X3']
mean_rep.reset_index(inplace = True, drop = True)
mean_rep
```

Out[10]:

	X1	X2	X3
0	5.0	3.0	4.0
1	5.0	3.0	4.0
2	5.0	3.0	4.0

In [11]:
```
#sample covariance matrix
covariance=(data-mean_rep).transpose().dot(data-mean_rep)/(len(data)-1)
covariance
```

Out[11]:

	X1	X2	X3
X1	13.0	-2.5	1.5
X2	-2.5	1.0	-1.5
X3	1.5	-1.5	3.0

Thus, the sample mean vector and the covariance matrix of x are:

$$\bar{x} = \begin{bmatrix} 5.0 & 3.0 & 4.0 \end{bmatrix} \qquad S = \begin{bmatrix} 13.0 & -2.5 & 1.5 \\ -2.5 & 1.0 & -1.5 \\ 1.5 & -1.5 & 3.0 \end{bmatrix}$$

(b) $c = \begin{bmatrix} 1 \\ 2 \\ -3 \end{bmatrix} \qquad x = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$ $\qquad$ Let $y = c'x = X_1 + 2X_2 - 3X_3$

Sample mean:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(x_{1i} + 2x_{2i} - 3x_{3i}) = \frac{1}{n}\sum_{i=1}^{n} x_{1i} + 2\left(\frac{1}{n}\sum_{i=1}^{n} x_{2i}\right) - 3\left(\frac{1}{n}\sum_{i=1}^{n} x_{3i}\right)$$

$$= \bar{x}_1 + 2\bar{x}_2 - 3\bar{x}_3$$

$$= 5 + 2(3) - 3(4)$$

$$= -1$$

Sample variances:

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2 = \frac{1}{3-1}\sum_{i=1}^{3}(y_i+1)^2 = \frac{1}{2}(1^2+7^2+6^2) = 43$$

$$\left(\because\ y_1 = 1+2(4)-3(3) = 0,\quad y_2 = 6+2(2)-3(6) = -8,\quad y_3 = 8+2(3)-3(3) = 5\right)$$

(c) $b = \begin{bmatrix}1\\1\\1\end{bmatrix}$ $x = \begin{bmatrix}X_1\\X_2\\X_3\end{bmatrix}$ Let $z = b'x = X_1+X_2+X_3$

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i = \frac{1}{n}\sum_{i=1}^{n}(x_{1i}+x_{2i}+x_{3i}) = \frac{1}{n}\left(\sum_{i=1}^{n}x_{1i}+\sum_{i=1}^{n}x_{2i}+\sum_{i=1}^{n}x_{3i}\right)$$

$$= \bar{x_1}+\bar{x_2}+\bar{x_3}$$

$$= 5+3+4$$

$$= 12$$

$z_1 = 1+4+3 = 8$

$z_2 = 6+2+6 = 14$

$z_3 = 8+3+3 = 14$

sample covariance:

$$\mathrm{Cov}(y,z) = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(z_i-\bar{z})}{n-1} = \frac{1}{2}\sum_{i=1}^{3}(y_i-\bar{y})(z_i-\bar{z})$$

$$= \frac{1}{2}\{(0+1)(8-12)+(-8+1)(14-12)+(5+1)(14-12)\}$$

$$= \frac{1}{2}(-4-14+12)$$

$$= -3$$

4. (a) Linear combinations of the components of $\underset{\sim}{X}$ are normally distributed.

If $\underset{\sim}{X} \sim N_3(\underset{\sim}{\mu}, \Sigma)$, then $\underset{\sim}{a}'\underset{\sim}{X} \sim N(\underset{\sim}{a}'\underset{\sim}{\mu}, \underset{\sim}{a}'\Sigma\underset{\sim}{a})$.

If $\underset{\sim}{a} = \begin{bmatrix}3\\-2\\1\end{bmatrix}$, $\underset{\sim}{a}'\underset{\sim}{X} = 3X_1-2X_2+X_3$

$$\underset{\sim}{a}'\underset{\sim}{\mu} = \begin{bmatrix}3 & -2 & 1\end{bmatrix}\begin{bmatrix}2\\-3\\1\end{bmatrix} = 6+6+1 = 13$$

$$\underset{\sim}{a}'\Sigma\underset{\sim}{a} = \begin{bmatrix}3 & -2 & 1\end{bmatrix}\begin{bmatrix}1 & 1 & 1\\1 & 3 & 2\\1 & 2 & 2\end{bmatrix}\begin{bmatrix}3\\-2\\1\end{bmatrix}$$

$$= \begin{bmatrix} 2 & -1 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ -2 \\ 1 \end{bmatrix} = 6 + 2 + 1 = 9$$

$\therefore\ 3X_1 - 2X_2 + X_3 \sim N(13, 3^2)$

(b) Let $\underset{\sim}{a} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$. Then $X_2 - \underset{\sim}{a}' \begin{bmatrix} X_1 \\ X_3 \end{bmatrix} = X_2 - a_1 X_1 - a_2 X_3$

$$\mathrm{Cov}(X_2, X_2 - a_1 X_1 - a_2 X_3) = \mathrm{Cov}(X_2, X_2) - a_1 \mathrm{Cov}(X_2, X_1) - a_2 \mathrm{Cov}(X_2, X_3)$$

($\because$ bilinearity)

$$= 3 - a_1 - 2a_2 \quad \text{(from covariance matrix } \Sigma)$$

Setting $3 - a_1 - 2a_2 = 0$, a possible solution is $a_1 = 1,\ a_2 = 1$

Thus, if $\underset{\sim}{a} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, X_2 and $X_2 - \underset{\sim}{a}' \begin{bmatrix} X_1 \\ X_3 \end{bmatrix}$ are independent.

(c) Conditional distributions of $\underset{\sim}{X}$ are multivariate normal.

$$\underset{\sim}{X} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \sim N_3(\underset{\sim}{\mu}, \Sigma) \text{ with } \underset{\sim}{\mu} = \begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix},\ \Sigma = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 2 \end{bmatrix}.$$

Let us rearrange and then partition as

$$\underset{\sim}{X} = \begin{bmatrix} X_3 \\ \cdots \\ X_2 \\ X_1 \end{bmatrix},\ \underset{\sim}{\mu} = \begin{bmatrix} 1 \\ \cdots \\ -3 \\ 2 \end{bmatrix},\ \Sigma = \left[\begin{array}{c|cc} 2 & 2 & 1 \\ \hline 2 & 3 & 1 \\ 1 & 1 & 1 \end{array}\right]$$

where $\underset{\sim}{\mu}_1 = 1$, $\underset{\sim}{\mu}_2 = \begin{bmatrix} -3 \\ 2 \end{bmatrix}$, $\Sigma_{11} = 2$, $\Sigma_{12} = \begin{bmatrix} 2 & 1 \end{bmatrix}$, $\Sigma_{21} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$, $\Sigma_{22} = \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}$

$$\Sigma_{22}^{-1} = \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}^{-1} = \frac{1}{2} \begin{bmatrix} 1 & -1 \\ -1 & 3 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{3}{2} \end{bmatrix}$$

Note that $|\Sigma_{11}| > 0$.

Then, for the conditional distribution $X_3 \mid X_1 = x_1,\ X_2 = x_2$, the mean is

$$\underset{\sim}{\mu}_1 + \Sigma_{12} \Sigma_{22}^{-1} (\underset{\sim}{x}_2 - \underset{\sim}{\mu}_2) = 1 + \begin{bmatrix} 2 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{3}{2} \end{bmatrix} \left(\begin{bmatrix} X_2 \\ X_1 \end{bmatrix} - \begin{bmatrix} -3 \\ 2 \end{bmatrix} \right)$$

$$= 1 + \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} X_2 + 3 \\ X_1 - 2 \end{bmatrix}$$

$$= 1 + \frac{1}{2}(X_1 + X_2 + 1)$$

$$= \frac{1}{2}X_1 + \frac{1}{2}X_2 + \frac{3}{2}$$

The covariance is

$$\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = 2 - \begin{bmatrix} 2 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{3}{2} \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

$$= 2 - \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

$$= 2 - \frac{3}{2}$$

$$= \frac{1}{2}$$

$$\therefore (X_3 \mid X_1 = x_1, X_2 = x_2) \sim N\left(\frac{1}{2}X_1 + \frac{1}{2}X_2 + \frac{3}{2}, \frac{1}{2}\right)$$