1

The following maximum likelihood estimates of the factor loadings for an $m = 1$ model were obtained:

Variable	Estimated factor loadings F_1
1. ln(length)	$.1022$
2. ln(width)	$.0752$
3. ln(height)	$.0765$

Using the estimated factor loadings, obtain the maximum likelihood estimates of each of the following.

(a) Specific variance

(b) Communalities

(c) Proportion of variance explained by the factor

(d) The residual matrix $S_n - \hat{L}\hat{L}^T - \hat{\Psi}$

透過定義我們知道

$$\widetilde{l}_{ij} = \sqrt{\hat{\lambda}_i}\hat{e}_{ij}$$

而 Communalities 則是

$$\widetilde{h}_i^2 = \sum_j \widetilde{l}_{ij}, \ \forall i$$

最後 Specific variance 則爲 Communalities 與 s_{ii} 之間的差異

$$\widetilde{\Psi} = s_{ii} - \widetilde{h}_i^2$$

已知

$$S = 10^{-3} \times \begin{bmatrix} 11.0720040 & 8.0191419 & 8.1596480 \\ 8.0191419 & 6.4167255 & 6.0052707 \\ 8.1596480 & 6.0052707 & 6.7727585 \end{bmatrix}$$

由於 maximum likelihood method 是使用 S_n 而非 S

$$S_n = \frac{n-1}{n} S = \frac{23}{24} S = 10^{-3} \times \begin{bmatrix} 10.6107 & 7.685 & 7.8197 \\ 7.685 & 6.1494 & 5.7351 \\ 7.8197 & 5.7351 & 6.4906 \end{bmatrix}$$

透過上述定義可以構建表格

Variable	Factor loadings F_1	Communalities $\widetilde{h}_i^2$	Specific variance $\widetilde{\Psi} = s_{ii} - \widetilde{h}_i^2$
1. ln(length)	.1022	0.0104	0.0001734
2. ln(width)	.0752	0.0057	0.0004941
3. ln(height)	.0765	0.0059	0.0006342
Cumulative Proportion	0.944		

根據表格資訊可得

(a) Specific variance 爲 $0.0001734, 0.0004941, 0.0006342$

(b) Communalities 爲 $0.0104, 0.0057, 0.0059$

(c) Proportion of variance explained by the factor 約爲 94.4%

(d) 而 residual matrix 則爲

$$S_n - \hat{L}\hat{L}^T - \hat{\Psi} == 10^{-6} \times \begin{bmatrix} 0 & 2.1673 & 1.4474 \\ 2.1673 & 0 & 0.112497 \\ 1.4474 & 0.112497 & 0 \end{bmatrix}$$

2

Perform a factor analysis of the data on bulls given in Table 1.10. Use the seven variables YrHgt, FtFrBody, PrctFFB, Frame, BkFat, SaleHt, and SaleWt. Factor the sample covariance matrix S and interpret the factors. Compute factor scores, and check for outliers. Repeat the analysis with the sample correlation matrix R. Compare the results obtained from S with the results from R. Does it make a difference if R, rather than S, is factored? Explain.

(a) Factor S using principal components (rotated)

(b) Factor S using maximum likelihood method (rotated)

(c) Factor R using principal components (rotated)

(d) Factor R using maximum likelihood method (rotated)

(e) Compare the results obtained from S with the results from R. Interpret the factors.

(f) Plot the scatter plots of factor2 vs factor1 in (a) and (c). Any outliers?

(a) 根據題意，使用 Python 計算可得:

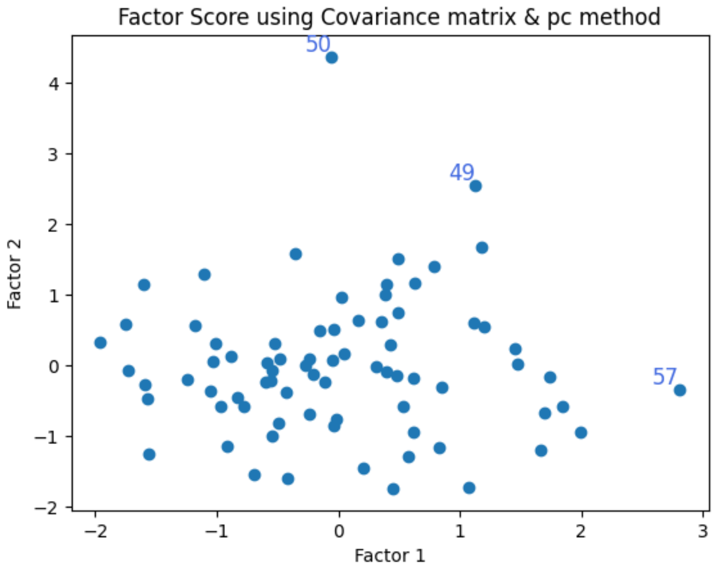
	Factor1	Factor2
YrHgt	0.62418	-0.02606
FtFrBody	0.99647	-0.08391
PrctFFB	0.67044	-0.28124
Frame	0.60612	-0.01173
BkFat	-0.13783	0.37510
SaleHt	0.71503	0.15422
SaleWt	0.62295	0.78226

Table 1: Principal component with two factors

	Factor1	Factor2	Factor3
YrHgt	0.50195	0.42460	0.32637
FtFrBody	0.25853	0.90600	0.33514
PrctFFB	0.83816	0.45576	0.18354
Frame	0.44716	0.42166	0.31943
BkFat	−0.60974	−0.06913	0.15478
SaleHt	0.40890	0.46689	0.50894
SaleWt	−0.13508	0.30219	0.94363

Table 2: Principal component with three factors

此外，使用 two factor 計算出的 factor score 我們可以畫出如下的 scatter plot，並以此我們可以判斷出第 49, 50 以及 57 號樣本為 outlier。



(b) 根據題意，使用 Python 計算可得到如下的 factor loading matrix：

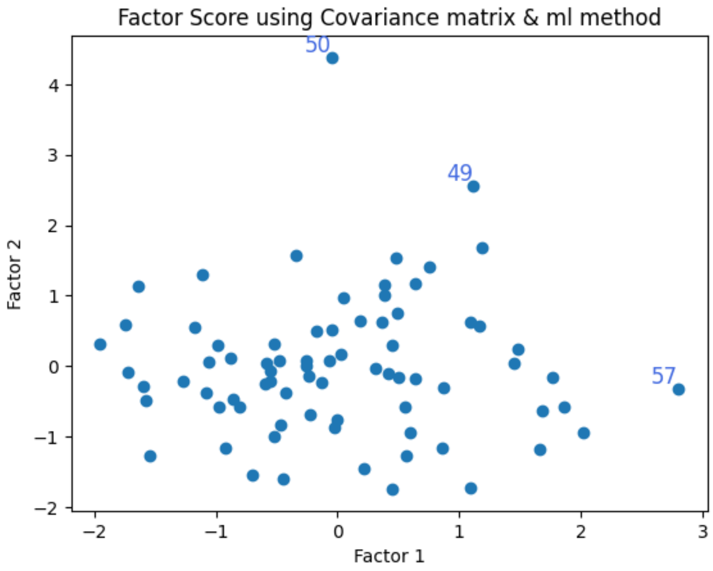
	Factor1	Factor2
YrHgt	0.920845	0.383505
FtFrBody	0.278508	0.957830
PrctFFB	0.303860	0.632094
Frame	0.862210	0.380402
BkFat	−0.339054	−0.077948
SaleHt	0.715102	0.522574
SaleWt	0.183127	0.526315

Table 3: Maximum likelihood with two factors

	Factor1	Factor2	Factor3
YrHgt	0.941498	0.285913	0.163903
FtFrBody	0.413567	0.505039	0.552527
PrctFFB	0.230785	0.946920	0.212326
Frame	0.890646	0.250911	0.180174
BkFat	−0.256189	−0.514123	0.272724
SaleHt	0.755024	0.269091	0.434350
SaleWt	0.253496	−0.050148	0.878884

Table 4: Maximum likelihood with three factors

此外，使用 two factor 計算出的 factor score 我們可以畫出如下的 scatter plot，並以此我們可以判斷出第 49, 50 以及 57 號樣本爲 outlier。



(c) 根據題意，使用 Python 計算可得到如下的 factor loading matrix：

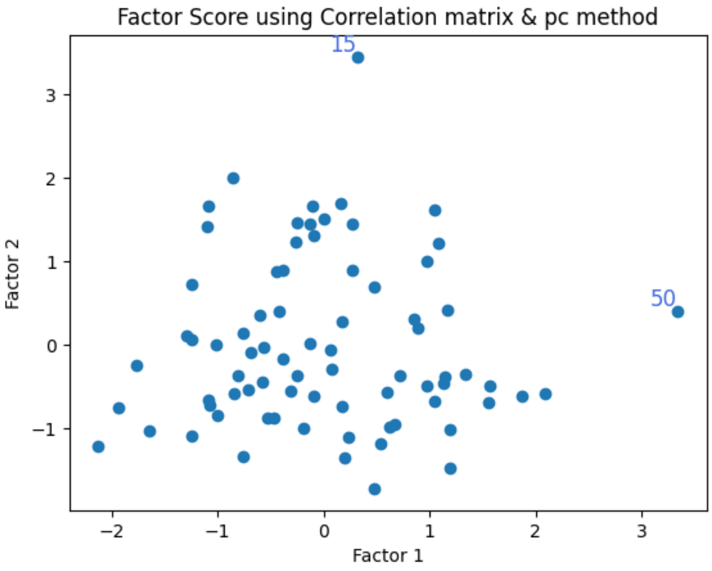
	Factor1	Factor2
YrHgt	0.873610	−0.271000
FtFrBody	0.848325	−0.058787
PrctFFB	0.610840	−0.530033
Frame	0.856428	−0.206434
BkFat	−0.165725	0.893988
SaleHt	0.920051	−0.110858
SaleWt	0.700950	0.539577

Table 5: Principal component with two factors

	Factor1	Factor2	Factor3
YrHgt	0.940960	0.269601	−0.081640
FtFrBody	0.447232	0.794499	0.205403
PrctFFB	0.261931	0.859017	−0.294890
Frame	0.937555	0.219088	−0.027825
BkFat	−0.230991	−0.339273	0.811990
SaleHt	0.833300	0.419203	0.108623
SaleWt	0.351752	0.430311	0.721907

Table 6: Principal component with three factors

此外，使用 two factor 計算出的 factor score 我們可以畫出如下的 scatter plot，並以此我們可以判斷出第 15 以及 50 號樣本爲 outlier。



(d) 根據題意，使用 Python 計算可得到如下的 factor loading matrix：

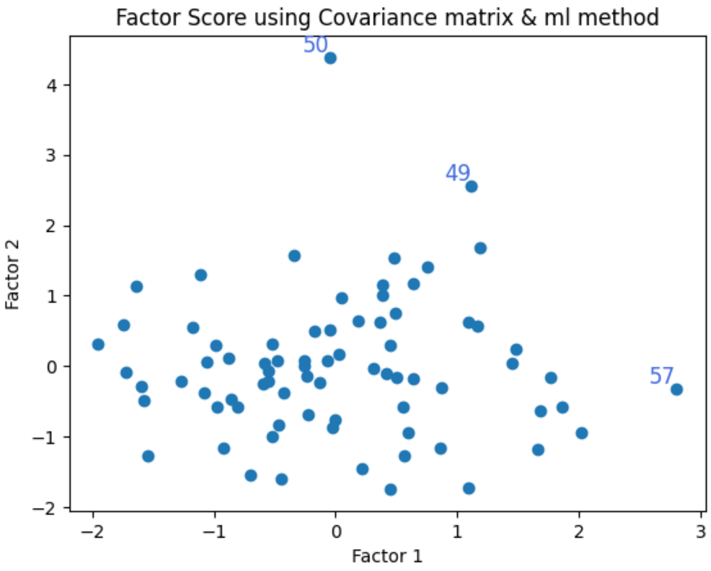
	Factor1	Factor2
YrHgt	0.920845	0.383505
FtFrBody	0.278508	0.957830
PrctFFB	0.303860	0.632094
Frame	0.862210	0.380402
BkFat	−0.339053	−0.077948
SaleHt	0.715102	0.522574
SaleWt	0.183127	0.526315

Table 7: Maximum likelihood with two factors

	Factor1	Factor2	Factor3
YrHgt	0.941498	0.285913	0.163903
FtFrBody	0.413567	0.505039	0.552527
PrctFFB	0.230785	0.946920	0.212326
Frame	0.890646	0.250911	0.180174
BkFat	−0.256189	−0.514123	0.272724
SaleHt	0.755024	0.269092	0.434350
SaleWt	0.253496	−0.050148	0.878885

Table 8: Maximum likelihood with three factors

此外，使用 two factor 計算出的 factor score 我們可以畫出如下的 scatter plot，並以此我們可以判斷出第 49, 50 以及 57 號樣本爲 outlier。



(e) 在使用 covariance matrix 以及 pc method 的條件下，根據每個 factor 對不同變數的 loading，我們可以發現 factor1 對 Yearling height at shoulder, Fatfree body,Frame and Sale Height at shoulder 有較高的正向效果，因此 factor1 可能代表了該牛隻的潛在體型上限；而 factor2 對 Sale weight 有較高的正向效果，對 percent of fatfree body 有負向效果，因此 factor2 可能代表了該牛隻肥胖程度。

在使用 correlation matrix 以及 pc method 的條件下，根據每個 factor 對不同變數的 loading，我們可以發現 factor1 對 Yearling height at shoulder, Frame and Sale Height at shoulder 有較高的正向效果，上述的幾個變數普遍與在各年齡段的身高有關，因此 factor1 可能代表了該牛隻父母的身高；而 factor2 對 Fat free body, Percent of fat-free body Sale height at shoulder and Sale weight 有較高的正向效果，因此 factor2 可能代表了該牛隻健碩程度。

根據以上描述，我們可以發現在使用 principal component 的前提下，使用 covariance matrix 與 correlation matrix 會計算出不同的 factor loading，其所代表的意涵也可能不同。

(f) 見 (a) (c) 小題的 scatter plot 與說明。

3

A survey is undertaken to determine consumer perceptions of six competing brands of soft drinks. The data collected are given in file drinks.DAT. The brands rated were as follows: (1) Pepsi; (2) Coke; (3) Gatorade; (4) Allsport; (5) Lipton tea; (6) Nestea. Respondents used a 7-point scale (1= strongly disagree to 7= strongly agree) to indicate their level of agreement/disagreement with the following 10 statements (in each of the statements substitute "Brand X" with the brands listed above.

- X_1: Brand X has a refreshing taste.
- X_2: I prefer Brand X because it has fewer calories than other drinks.
- X_3: Brand X quenches my thirst immediately.
- X_4: I like the sweet taste of Brand X.
- X_5: I prefer drinking Brand X after workouts and sports because it gives me energy.
- X_6: I prefer Brand X because it comes in environment friendly packaging.
- X_7: Brand X has minerals and vitamins that help quench my deep down body thirst.
- X_8: Brand X has a unique flavor of its own.
- X_9: Brand X has the right mix of minerals and vitamins that are healthy for my body.
- X_{10}: I prefer to drink Brand X when I am really thirsty.

Use principal components factor analysis to analyze the data (correlation matrix)

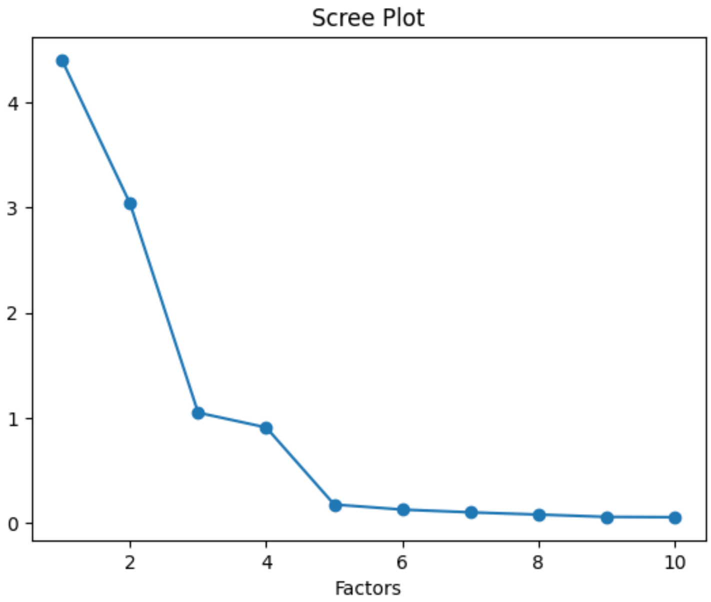
(a) Determine the appropriate number of factors to effectively account for the variance in the data. Show the Scree plot and rotated factor loadings.

(b) Label the factors and explain their meanings.

(c) Calculate the ”average” factor scores of each brand. Then make all scatter plots of any two factor scores. Note that you will have only six points (brands) in each scatter plot.

(d) Use the scatter plots to interpret the positions of the six brands.

(a) 根據題意，我們可以畫出相應的 scree plot 並列出所有 eigenvalues 的值如下；



λ_i	Eigenvalues
λ_1	4.4400076
λ_2	3.037141
λ_3	1.049598
λ_4	0.908929
λ_5	0.176757
λ_6	0.128433
λ_7	0.102301
λ_8	0.081238
λ_9	0.059125
λ_{10}	0.056402

根據上表，我們可以得知有三個 eigenvalues 的值大於一，故我們取三個 factor 去進行 factor analysis。經計算後我們可得到 factor loading matrix 如下；

Variable	Factor1	Factor2	Factor3
X_1	-0.235599	0.877814	0.158810
X_2	0.396774	0.160685	0.829289
X_3	0.861227	-0.214173	0.273559
X_4	-0.114610	0.925119	0.022003
X_5	0.411827	0.194144	0.841072
X_6	0.325409	0.323635	-0.585740
X_7	0.873762	-0.282214	0.207741
X_8	-0.225925	0.923505	0.047481
X_9	0.438543	0.165765	0.827412
X_{10}	0.866995	-0.240396	0.240751

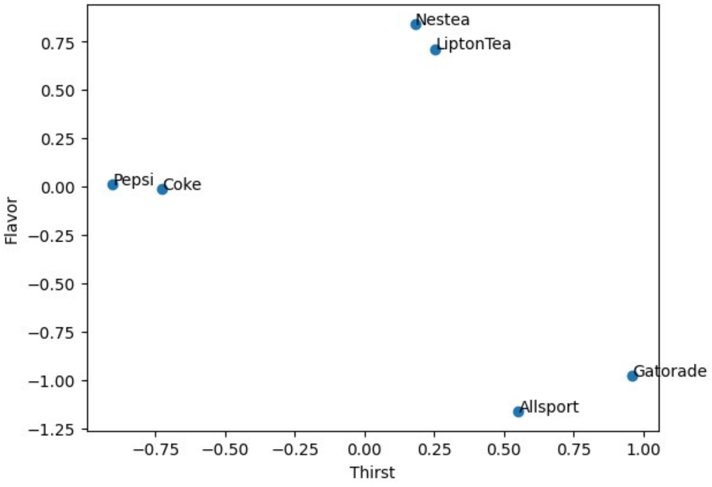
(b) 由上方裝有 factor loading 的 dataframe 我們可以發現 factor1 對 X_3, X_7 以及 X_{10} 有較強的正向效果，這三個變數都與飲料本身的解渴能力有關，因此 factor1 可能代表了該飲料的解渴能力；而 factor2 對 X_1, X_4 以及 X_8 有較強的正向效果，這三個變數都與飲料的口味有關，因此 factor2 有可能代表了消費者對飲料口味的偏好程度；factor3 對 X_2, X_5 以及 X_9 有較強的正向

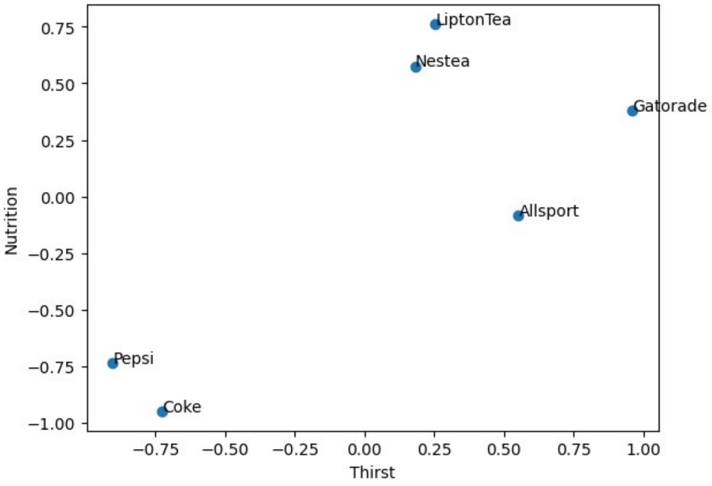
效果，這三個變數都與飲料是否能帶來足夠的營養與能量有關，因此 factor3 可能代表了該款飲料的健康程度。

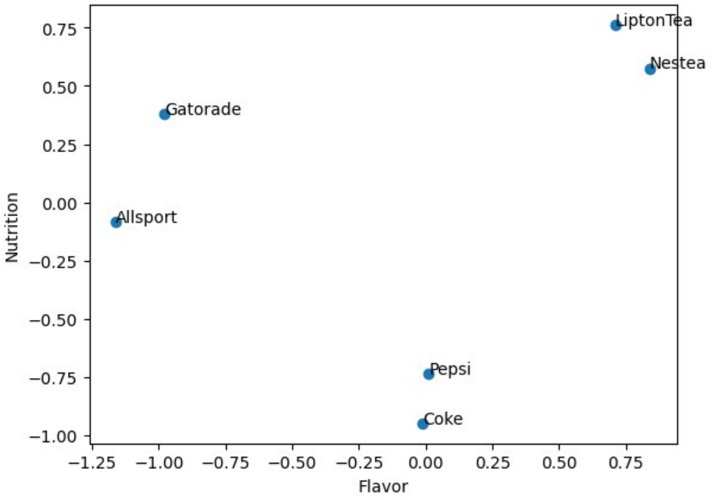
(c) 計算出資料的 factor score 後依照品牌別計算出 average factor score 如下，同時畫出任兩個 factor score 組成的 scatter plot。其中

$$(\text{Factor1}, \text{Factor2}, \text{Factor3}) = (\text{Thirst}, \text{Flavor}, \text{Nutrition})$$

Beverage	Thirst	Flavor	Nutrition
Pepsi	−0.903023	0.011078	−0.734029
Coke	−0.724844	−0.010301	−0.948358
Gatorade	0.960126	−0.977955	0.378326
Allsport	0.552470	−1.160611	−0.084057
LiptonTea	0.254755	0.711510	0.762124
Nestea	0.182522	0.840664	0.575364


Nestea
LiptonTea
Pepsi
Coke
Gatorade
Allsport
Flavor
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.25
Thirst
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00


LiptonTea
Nestea
Gatorade
Allsport
Pepsi
Coke
Nutrition
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Thirst
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00


LiptonTea
Nestea
Gatorade
Allsport
Pepsi
Coke
Nutrition
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Flavor
−1.25
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75

(d) 在所有飲料品牌中，消費者認爲 Pepsi 與 Coke 在解渴方面的表現較差，而 Gaatorade 與 Allsport 在這方面的表現較好；而在味道偏好方面，消費者相對篇好 LiptonTea 與 Nestea，相對厭惡 Gatorade 以及 Allsport；最後，在營養價值與能量供給方面，消費者同樣較爲青睞 LiptonTea 與 Nestea，而不喜歡 Pepsi 與 Coke。同時我們可以發現各品牌在 Thirst 與 Nutrition 兩個 Factor 的表現上呈現一定的正相關。