Name: Tan Wen Tao Bryan

Class: DAAA/FT/2A/01

Student ID: 2214449

**Qn 1**

**a)**

**-** Use correlation matrix instead of covariance matrix as the values of the data are of different units and magnitude

**PCA**

A picture containing text, screenshot, font, number

Description automatically generated

PC1 explains 46.67% of the total variance. PC2 explains 26.78% of the total variance. PC3 explains 11.34% of the total variance.

**Screeplot**

A picture containing line, plot, screenshot, text

Description automatically generated

**Extraction of PCs**

Numbers of PCs to extract:

* By Kaiser’s rule, extract the first 2PC as whose eigenvalues (3.73 & 2.14) respectively are >1.
* 1st 3 PC s already accounted for 84.79% of total variance, PC3 account for 11.34% of total variance, which could be too high to discard.
* Screeplot shows elbow at PC4, suggesting 3PC to extract.

Let’s extract the first 3 PCs only.

**Loading Plot**

A picture containing text, screenshot, diagram, line

Description automatically generatedA picture containing text, line, screenshot, diagram

Description automatically generatedA picture containing text, screenshot, diagram, line

Description automatically generated

**Interpretation of PCs:**

PC1:

ŷ1 = 0.4731z1 + 0.2974z2 + 0.0517z3 + 0.0786z4 + 0.3650z5 + 0.3601z6 + 0.4411z7 + 0.4706z8

The loadings on carb and sugar are quite small, whereas the loadings on protein, fat, Vitamin A, Vitamin B6, Vitamin B12 and calcium are bigger. This PC seems to measure the average nutrients composition in dairy products.

Protein has the highest loading, so z1 has the greatest impact in value of ŷ1 for PC1.

PC2:

ŷ2 = -0.1048z1 – 0.3213z2 + 0.6244z3 + 0.6061z4 – 0.1748z5 + 0.2503z6 + 0.1712z7 – 0.0778z8

The loadings on Vitamin B6 and Vitamin B12 are quite small compared to loadings on Carb and Sugar. The loadings on Calcium, Protein, Vitamin A and Fat are opposite in sign to the loadings on Vitamin B6, B12, Carb and Sugar. This PC seems to measure a contrast of nutrients that are involved in energy metabolism (carbs, sugar, Vitamin B6, B12) against nutrients that are involved in physiological (Calcium and Protein) and dietary fats (Vitamin A, Fat) related components in dairy products.

Sugar has the highest loading, so z4 has the greatest impact of ŷ2 for PC2.

PC3:

ŷ3 = 0.1550z1 – 0.5709z2 – 0.2856z3 – 0.3362z4 – 0.4771z5 + 0.3537z6 + 0.3014z7 + 0.1095z8

The loadings on Protein, Vitamin B6, Vitamin B12, Calcium are opposite in sign to the loadings on Fat, Carb, Sugar and Vitamin A. This PC seems to measure a contrast of beneficial nutrients (Protein, Vitamin B6, Vitamin B12, Calcium) against the nutrients that are of less indicative of overall nutritional quality (Fat, Carb, Sugar and Vitamin A) in dairy products when taking excessive amounts.

Fat has the highest loading, so z2 has the greatest impact of ŷ3 for PC3.

**Score Plot**

A picture containing text, screenshot, diagram, plot

Description automatically generatedA picture containing text, screenshot, diagram, line

Description automatically generatedA picture containing text, screenshot, diagram, plot

Description automatically generated

**Interpretations of score plots:**

From the score plot, most of the yoghurt, ice cream, milk and cream tend to score low on PC1.

Most of the cheese and cream tend to score low while some cream and milk tend to score high as outliers on PC2. PC2 also has most of the milk and yoghurt scoring near zero on PC2. Most yoghurt, cream, cheese and milk score near zero on PC3.

**b)**

**i)** Using PC1, some milk and some cheeses are low in carbs and sugar but high in other nutrients.

**ii)** Using PC2, some milk and some creams are high in carbs and sugar but low in other nutrients.

**c)**

**Score Plots with Product1 & Product2 markers**

**A picture containing text, screenshot, diagram, plot

Description automatically generatedA picture containing text, screenshot, diagram, line

Description automatically generatedA picture containing text, screenshot, diagram, plot

Description automatically generated**

Using PC1, Product 1 is Cheese

Using PC2, Product 2 is Yoghurt

**d)**

In part(a), by conducting a PCA analysis on the dairy products and creation of the loading plots, I see that different types of nutrients have different variation on the different PCs. For example, carbs and sugar scoring low while other nutrients scoring high on PC1. Through this, I understand the relationship between the variables (nutrients) and the PCs, and we can interpret what each of the PC shows about the different nutrients. For example, PC1 seems to measure the average nutrients composition in dairy products.

In part(c), based on 2 given products provided nutritional values, I managed to locate them on the score plots of the different dairy products plotted based on their PCs. By looking at the score plot, I can identify whether the products are similar to or different from other products in terms of nutritional composition.

As such, I would have expected that products of the same type like milk to cluster together, while products of different types like cheese and yoghurt to separate from one another on the same PC. However, from plotting the score plot, I have realized in actual fact that is not the case, even products with the same type are deviating from one another while there are products with different type are clustering together on the same PC. This is because even products of the same type are made up of different nutritional composition.

I would also have expected the PCs to identify the products that exist as outliers and position them away from the main clusters of observations. However, in actual fact, only if the outliers are clearly distinct and obvious, then it would be positioned away from the main clusters of observations.

**Qn 2**

**a)**

**-** Use correlation matrix instead of covariance matrix as the values of the data are of different units and magnitude

**PCA**

**A screenshot of a graph

Description automatically generated with low confidence**

PC1 explains 61.69% of the total variance. PC2 explains 17.2% of the total variance. PC3 explains 9.14% of the total variance.

**Screeplot**

A picture containing text, plot, line, screenshot

Description automatically generated

**Extraction of PCs**

Numbers of PCs to extract:

* By Kaiser’s rule, extract the first 3PC as whose eigenvalues (7.40, 2.06 & 1.10) respectively are >1.
* 1st 3 PC s already accounted for 88.03% of total variance, PC3 accounts for 9.14% of total variance, which could be too high to discard.
* Screeplot shows elbow at PC4, suggesting 3PC to extract.

Let’s extract the first 3 PCs only.

**Loading Plot**

A picture containing text, screenshot, line, diagram

Description automatically generatedA picture containing text, screenshot, diagram, plot

Description automatically generatedA picture containing text, screenshot, diagram, line

Description automatically generated

**Interpretation of PCs:**

PC1:

ŷ1 = 0.2750z1 + 0.3203z2 + 0.3211z3 + 0.1069z4 + 0.3512z5 + 0.2651z6 + 0.3315z7 + 0.0044z8 + 0.3065z9 – 0.3127z10 – 0.3222z11 + 0.3287z12

The loadings on Compression-Ratio and Height are quite small with Compression-Ratio nearly hitting 0, whereas the loadings on Wheel-base, Length, Width, Curb-weight, Cylinders, Engine, Horsepower, Price are bigger. The loadings on Highway and City are opposite in sign to the loadings on Compression-Ratio, Height, Wheel-base, Length, Width, Curb-weight, Cylinders, Engine, Horsepower and Price. This PC seems to measure the relationship between the attributes of the cars, whether they have a direct or inverse relationship.

Curb-weight has the highest loading so it has the greatest impact on ŷ1 for PC1.

PC2:

ŷ2 = -0.3882z1 – 0.2562z2 – 0.1903z3 – 0.5242z4 – 0.0935z5 + 0.2260z6 + 0.1345z7 – 0.4762z8 + 0.3046z9 – 0.2087z10 – 0.1577z11 + 0.0821z12

The loadings on Wheel-base, Length, Width, Height, Curb-weight, Compression-ratio, City and Highway are in opposite sign to the loadings on Price, Engine, Cylinders & Horsepower. This PC seems to measure a contrast of the attributes related to the physical characteristics and fuel efficiency of the cars against the attributes related to the performance and price of the cars.

Height has the highest loading so it has the greatest impact on ŷ2 for PC2.

PC3:

ŷ3 = -0.1284z1 – 0.1378z2 + 0.0434z3 – 0.3313z4 + 0.0387z5 + 0.3976z6 + 0.2672z7 + 0.6076z8 + 0.0419z9 + 0.3222z10 + 0.3045z11 + 0.2334z12

The loadings on Width, Curb-Weight and Horsepower are quite small whereas the loadings on Engine, Price, Cylinders, Highway, City and Compression-ratio are bigger. The loadings on Wheel-base, Length & Height of the cars are opposite in sign to Width, Curb-Weight, Horsepower, Engine, Price, Cylinders, Highway, City and Compression-ratio. This PC seems to measure a contrast of the cars’ physical dimensions against attributes that are involved in the cars’ fuel efficiency and price.

Compression-ratio has the highest loading so it has the greatest impact on ŷ3 for PC3.

**Score Plot**

A picture containing text, screenshot, diagram, plot

Description automatically generatedA picture containing text, screenshot, diagram, line

Description automatically generatedA picture containing text, screenshot, diagram, plot

Description automatically generated

**Interpretations of score plots:**

From the score plot, most of hatchback cars tend to score low on PC1. Most sedan and wagon cars tend to score low on PC2. Most wagon cars tend to score low, while most sedan and hatchback cars tend to score near zero on PC3.

**b)**

PCA results of this dataset are not as clear or obvious as the dairy nutrition dataset in Question 1. PCA is not as effective on this dataset as it is to the dairy nutrition dataset in Qn1 as the clusters formed in the score plot is not as obvious. Moreover, car types like sedan and hardtop do not have sufficiently large sample size, causing the correlation coefficient to be less reliable, limiting the effectiveness of PCA. PCA is useful for dimension reduction, where it transforms a high dimensional dataset to a low dimensional space while retaining as much information as possible, making classification or clustering tasks much easier. PCA can also be used for feature extraction by identifying the important features or patterns that contribute to classification and clustering of data. Moreover, PCA allows for a visualization and interpretation of high-dimensional data in allow dimensional space. For example, the score plot makes it easier to observe patterns, clusters, or separations. However, PCA is limited as PCA assumes linearity and does not reflect non-linearity relationship so if linearity fails, PCA is not effective.