Protein Composition Analysis

Statistical analysis by Alex Shomali

Protein Composition Analysis

Introduction: Proteins are an essential component in any healthy diet and responsible for a vast amount of different types of bodily and cognitive functions, such as the repair of muscles and the construction of hormones. Proteins are composed of amino acids, often referred to as the "building blocks" of the body. A diet of animal based proteins such as meat and fish contain every amino acid needed for bodily and cognitive operation, while plant based proteins such as beans and vegetables often lack some of the required essential amino acids. In addition, high protein diets are associated with health issues such as heart disease and cancer. Data from the USDA National Nutrient Database for Standard Reference contains 1244 observations of composition of different protein sources. Understanding the compositions will be useful to people with plant based diets, such as vegetarians or vegans so they will be able to attain all the essential amino acids on a plant based diet and reduce health risks.

Objectives: Utilise unsupervised dimensionality reduction techniques to determine and understand the nutrients composition of plant based protein sources and animal based protein sources, as well as determine the underlying structures within the nutrients compositions.

Methodology: Principal component analysis (PCA) was used to reduce dimensions and understand the nutrients composition between plant based and animal based protein sources. Factor analysis (FA) was used to determine the underlying structures within the nutrients compositions. To test assumptions, the Hene-Zeker multivariate normality test was used and Chi-squared plots were used to detect multivariate outliers.

Statistical Analysis: The nutrients considered in the analysis were niacin, phosphorus, fat, copper, protein, folate, vitamin B6, thiamine, zinc, vitamin C, magnesium, selenium, vitamin A, vitamin B12 and iron. The protein sources considered were plant based sources and animal based sources. The statistical software package R, has been used to generate table data and figures for analysis. For PCA, the R library "factoextra" was used to generate improved visualisation and for FA, the libraries "psych" and "nFactors" were utilised. In addition, the library "MVN" was used to conduct multivariate normality tests.

Results: Below are the results from the conducted PCA. It is ideal that the data is reduced such that roughly 80% of the variability in the data is explained.

Principle Component	**Proportion of explained variance**	**Cumulative proportion explained**
Principle Component 1	26.9%	26.9%
Principle Component 2	13.3%	40.2%
Principle Component 3	11.2%	51.4%
Principle Component 4	8.5%	59.9%
Principle Component 5	7.4%	67.3%
Principle Component 6	6.6%	73.9%
Principle Component 7	6.0%	79.9%
Principle Component 8	5.8%	85.7%
Principle Component 9	3.2%	88.9%
Principle Component 10	3.0%	91.9%
Principle Component 11	2.6%	94.5%
Principle Component 12	2.1%	96.6%
Principle Component 13	1.5%	98.1%
Principle Component 14	1.1%	99.2%
Principle Component 15	0.8%	100%

Table 1.0: Variance explained by principle components

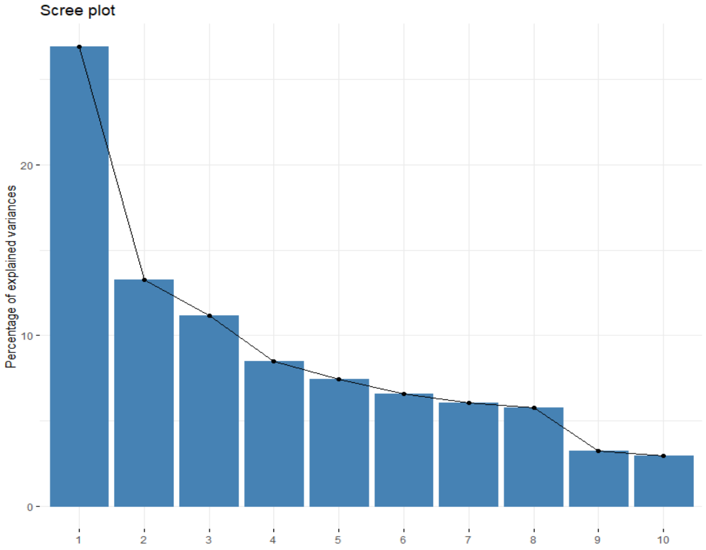


Figure 1.1: Scree plot for first 10 dimensions

From Figure 1.1, it is evident that the "elbow" of the scree plot is at the third dimension. That is, after the third dimension, the variance explained by all subsequent dimensions is very minimal. However, from Figure 1.0, the cumulative proportion of variance explained by the first three dimensions is only 51.4%. It is ideal that at least roughly 80% of the variability is explained by the dimensions. Thus, the ideal number of principle components is seven as the cumulative proportion of variance explained the seventh dimension is 79.9%.

The loadings on variables for the optimal number of principle components are shown in the below table.

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Niacin	0.191	0.474	0.181	0.213	0.220	-	-
Phosphorus	0.267	0.179	-0.282	-0.273	-0.291	-	-0.165
Fat	-	0.167	-0.186	-0.171	-0.118	0.870	-
Copper	0.355	-0.315	-	-0.155	0.277	-	-
Protein	0.303	0.403	-	-0.210	-0.124	-0.176	-
Folate	0.349	-0.236	-0.203	0.259	-0.244	-	0.159
Vitamin B6	0.276	0.391	0.132	0.307	0.184	-0.102	-
Thiamin	-	-	-0.107	0.453	0.367	0.270	-0.262
Zinc	0.193	-0.207	-	-0.439	0.553	-	-0.208
Vitamin C	-	-0.169	0.186	0.179	-0.306	-	-0.849
Magnesium	0.304	-	-0.490	-	-	-0.162	-
Selenium	0.223	0.214	0.385	-0.352	-0.177	-0.108	-
Vitamin A	0.253	-0.224	0.347	0.200	-0.305	0.252	0.309
Vitamin B12	0.247	-0.172	0.472	-	-	0.101	-
Iron	0.394	-0.211	-	0.147	-	-	-

Table 1.2: Loadings on each principle component

The first principle component has a positive correlation with all the variables, excluding fat, thiamine and vitamin C, to which there is no loading. The relationship between these nutrients and PC1 is weak. In contrast to the other nutrients, iron, copper and folate load highly on PC1. In contrast to PC1, PC2 has a negative correlation with most nutrients. However, it has a positive correlation with Niacin, phosphorus, fat, protein and selenium with no loadings on thiamine and magnesium. The relationship is once again weak for the nutrients, besides niacin and protein which are considered strong when contrasted against the other variables. PC2 loads highly on niacin and vitamin B6 which is opposite to that of the loadings on PC1. The third PC has no loadings on copper, protein, vitamin C and iron. It has a positive correlation with selenium, vitamin A and vitamin B12 and also loads highly on these nutrients. It also has a strong (relative to other variables) negative loading on magnesium, the other variables only have a light loading. PC4 has a positive correlation with iron, vitamin A, vitamin C, thiamin, vitamin B6, folate and niacin and a negative correlation with the other nutrients; excluding vitamin B12 and magnesium to which there is no loading. There is a strong positive correlation (contrasted against other variables) for thiamin and a strong negative loading for zinc (when contrasted against the other nutrients). PC5 is very similar to PC4; however, there is no loading on zinc and copper has a positive correlation to PC4, while vitamin C and folate are negative loadings. In contrast to PC4, PC5 has a strong positive correlation with zinc. PC6 does not load on many nutrients; it has no loading on niacin, phosphorus, copper, zinc, vitamin C and iron. All the correlations on the remaining nutrients are weak. However, there is an extremely high positive correlation with fat, especially contrasted against other nutrients in all other PCs. Much like PC6, PC7 has no loading on many nutrients. The remaining nutrients have a weak positive or negative correlation with PC7. However, vitamin C has a very strong negative correlation with PC7 as is the only variable that has a high loading.

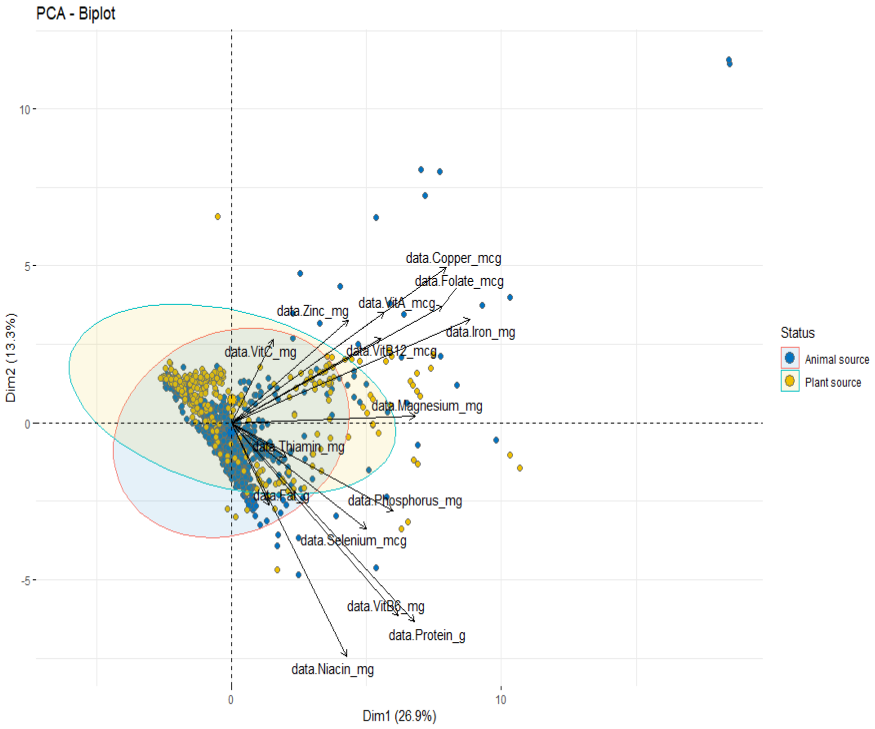


Figure 1.3: Biplot of animal and plant based sources

Figure 1.3 explanation: The biplot can be used to understand relationships among variables and nutrients compositions for different protein sources. It is evident the nutrients composition for animal sources are vitamin C, zinc, vitamin B12, vitamin A, folate, copper and iron. These variables are all positively correlated with eachother. For plant based sources, the composition is thiamine, phosphorus, fat, selenium, niacin, vitamin B6 and protein. These variables are also positively correlated with each other. The nutrients in the two composition clusters have a correlation; however, it is a weak correlation as the angle is approaching 90 degrees. Magnesium is in the composition of both dietary sources and it is mildly correlated with both of the two sets of nutrients compositions. In animal sources, the most significant nutrients are copper, folate and iron while vitamin C and zinc are not as significant. For plant based sources, the length of the vectors indicate that niacin, protein and vitamin B6 are the most significant nutrients. thiamin, fat, phosphorus and selenium are not as significant.

After principle component analysis, factor analysis was utilised to determine the underlying structures within the nutrients. To determine the optimal number of factors, a parallel analysis was utilised. The results are represented in the figure below.

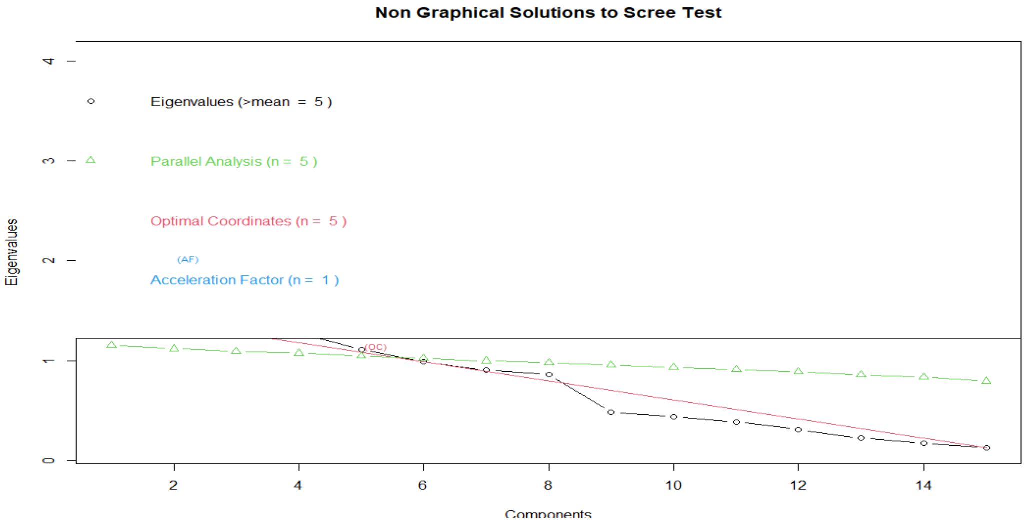


Figure 2.0: Parallel analysis

Figure 2.0 shows a parallel analysis that was conducted to determine the number of underlying structures. From the graph, it is easy to see that the optimal number of factors in five. Using this information, a factor analysis was able to be conducted to determine the underlying factors of the nutrients.

Using the number of factors as 5, a factor analysis was conducted. To determine the coefficients, the maximum likelihood estimation approach was utilised. In addition, a varimax rotation was used to assist in assigning nutrients to specific factors. The results from the factor analysis are shown below.

Variable	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 4**	**Factor 5**
Folate	0.818	-	0.466	-	-
Magnesium	0.794	-	-	-	-
Iron	0.509	-	0.406	-	-
Niacin	-	0.991	-	-	-
Vitamin B6	-	0.689	-	-	-
Vitamin A	-	-	0.829	-	-
Vitamin B12	-	-	0.56	-	-
Phosphorus	-	-	-	0.605	-
Protein	-	-	-	0.789	-
Selenium	-	-	-	0.590	-
Copper	0.409	-	-	-	0.877
Zinc	-	-	-	-	0.636
Fat	-	-	-	-	-
Thiamin	-	-	-	-	-
Vitamin C	-	-	-	-	-

Table 2.1: Factor analysis on nutrients

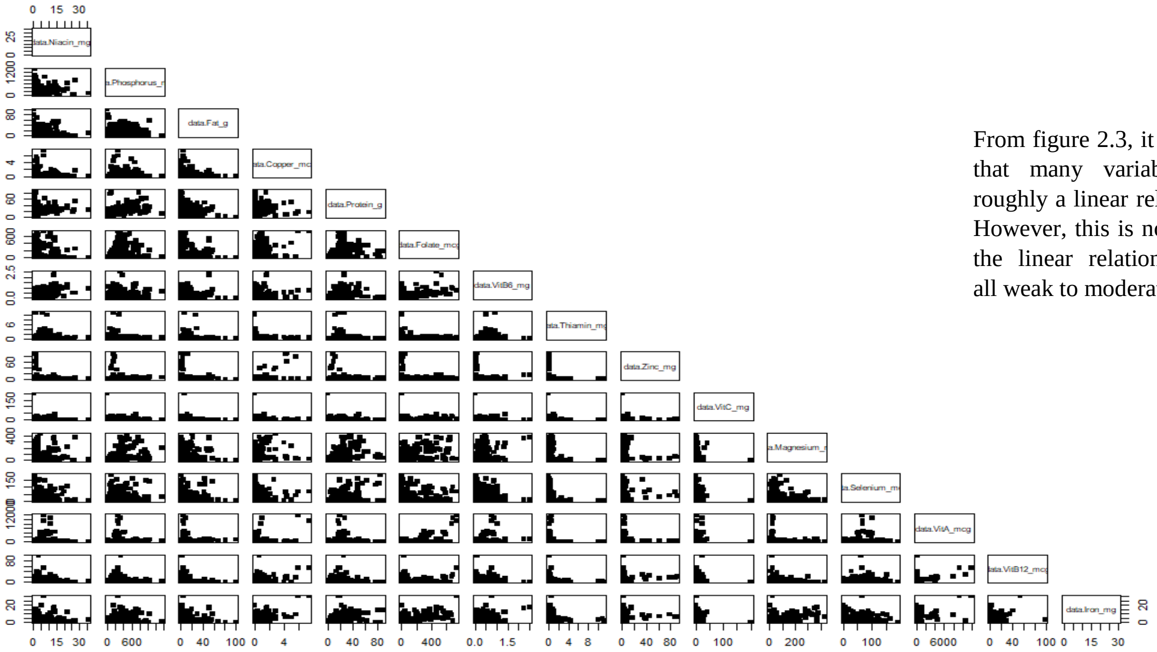
Using a cut off of 0.4, each of the nutrients can be assigned to a factor. Factor 1 consists of folate, magnesium and iron. Factor 2 consists of niacin and vitamin B6. Factor 3 consists of vitamin A and vitamin B12. Factor 4 consists of phosphorus, protein and selenium. Factor 5 consists of copper and zinc. Fat, thiamin and vitamin C did not meet the 0.4 cut off and have not been assigned a factor. For appropriate naming of factors, it would be best to consult someone in the biological or chemical sciences.

	Factor 1	**Factor 2**	**Factor 3**	**Factor 4**	**Factor 5**
Proportion explained variance	13.2%	11.4%	10.9%	10.3%	10.0%
Cumulative explained variance	13.2%	24.5%	35.5%	35.8%	55.8%

Table 2.2: Variance explained by factors

The variance explained by the factors is not ideal. Each factor explains roughly 10% of the variability in data, with the highest being factor 1 at 13.2% which drops off to 10% by factor 5. The cumulative variance of all the extracted factors is only 55.8%, thus, it is evident there is heavy information loss present.

The assumptions of factor analysis were also tested. The first of these assumptions is linearity between variables. To test this, the matrix scatter plot was used to observe the relationship between variables.

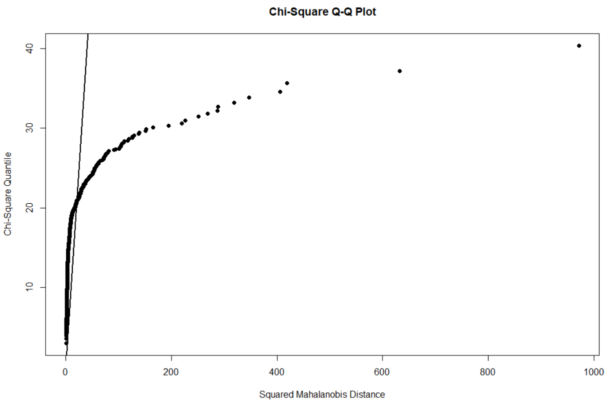


From figure 2.3, it is evident that many variables have roughly a linear relationship. However, this is not ideal as the linear relationships are all weak to moderate.

Figure 2.3: Matrix scatterplot

Factor analysis is very sensitive to multivariate outliers. To detect these, a chi-squared plot was used.

Figure 2.4: chi-squared plot



From figure 2.4, no immediate outliers are apparent. However, one observation has a squared Mahalanobis distance of nearly 1000 which could be considered a multivariate outlier in a conservative setting.

Another assumption is adequate sample size. It is ideal to have 20 observations to one variable. In the data set, there was 1244 observations across 15 variables, thus, the assumption of adequate sample size is met. While it is not a mandatory assumption, having multivariate normality in the data set produces significantly enhanced results. To test multivariate normality, Henze-Zirkler multivariate normality test was utilised. It was concluded that multivariate normality could not be assumed, $p < 0.01$ and HZ = 50.38. Shapiro Wilk's test was used on each of the variables and revealed that all nutrients were not univariate normally distributed as $p < 0.01$ for all 15 separate tests.

The final assumption is multicollinearity. To test this, each individual variable was regressed onto the remaining set of variables. It is ideal that the R squared value is not high, but at the same time, not low. The results are presented below.

Variable	**R squared**
Niacin	0.603
Phosphorus	0.494
Fat	0.146
Copper	0.701
Protein	0.646
Folate	0.713
Vitamin B6	0.603
Thiamin	0.106
Zinc	0.470
Vitamin C	0.056
Magnesium	0.681
Selenium	0.473
Vitamin A	0.671
Vitamin B12	0.472
Iron	0.653

Table 2.5: multicollinearity analysis

From table 2.5, it is evident that most of the nutrients meet this assumption. However, fat, vitamin C and thiamin have a very low R squared and do not meet the assumption which led to difficulty in assigning these nutrients to a factor during factor analysis.

Discussion: While this experiment has yielded results that have practical implications and can be useful, there is room for improvement. To better understand the nutrients compositions, instead of grouping the data into “animal sources” and “plant sources”, the data could have been collected to provide insight towards the type of animal and plant based sources. For example, plant based sources could have been broken down into “vegetable sources”, “bean sources” and “nut sources”. This would have allowed for better insight towards the composition of certain sub-groups of food rather than a very broad overview and perhaps provide insight towards naming the five factors. As a dimensionality reduction technique, the results for PCA are not ideal as the dimensions had only been reduced by half to explain 80% of the variability in the data. The data was also not multivariate normally distributed and the linear relationships were not very strong in the data set. In addition, some nutrients had a very low level of multicollinearity. This has led to FA results which are not ideal and a high level on information loss. Regardless, this study has a range of practical implications. With the composition of plant and animal sources understood, people on restrictive plant based diets can be advised on the missing nutrients needed from an animal based diet which has the potential to reduce health risk among the vegetarian and vegan sub culture. In addition, now understanding the composition of animal based sources can help reduce risks. As high animal source diets are known to have associations with diseases such as heart failure and cancer, understanding the nutritional composition can aid in understanding which nutrients to reduce an intake of. This will aid in the reduction of health risks associated with animal source based diets.

Conclusion: This study has found that the composition of animal based sources and plant based sources differ tremendously. In particular, the nutrients most significant to animal sources were vitamin C, zinc, vitamin B12, vitamin A, folate, copper and iron. While for plant based sources, the most significant nutrients were thiamine, phosphorus, fat, selenium, niacin, vitamin B6 and protein. Furthermore, five underlying structures were found in the nutrients which could be used to provide insightful information on which nutrients are vital to certain diets. While the dimensionality reductions for PCA were not ideal and FA had a significant level of information loss, the findings of this study can still be very helpful and used to significantly reduce health risks associated with plant based diets and reduce health risks such as cancer and heart disease that are associated with animal based diets. This study is also a great stepping stone for further analysis within the health sciences, particularly for understanding certain diets and food compositions.