An investigation into the USDA Data Set

Joseph Maliszewski

March 2020

1 Introduction

The data set investigated was the USDA National Nutrient DB. It was created by the US Department of Agriculture, (USDA) which hosts an integrated data system that provides expanded nutrient profile data and links to related agricultural and experimental research.[1]. The database allows for free use of this data to be used for academic and commercial purposes. The database consists of 23 different nutrients which are the columns, and 8618 food samples which are the rows. This data set therefore has 23 dimensions. The data already provided categorising of its own, which separated the samples into 25 different food groups. Examples include dairy, fats and oils, beverages and beef products. A more cleaned and validated version of the data base was provided by Craig Kelly at data.world.[2]

2 Hypotheses

1. Food groups in the data set would present a unique composite of nutrient ratios per food group.
2. Foods with high fat content and high carbohydrate content would be indicative of high levels of energy.
3. It would be possible to accurately predict Energy (kcal) exclusively from the other 22 of the 23 nutrients (dimensions) into the class of low, medium, high, and very high energy levels.

3 Method

All investigations in order to answer these hypotheses were undertaken with python. Libraries utilized were sklearn, numpy, pandas, matlibplot and seaborn. The statistical methods used to investigate the data were PCA, LDA and tSNE, all three being dimensionality reduction techniques, and further arbitrarily chosen 2D projections of the original dimensions.

3.1 Usefulness of dimensionality reduction

The two primary reasons for using dimensionality reduction are for data exploration, and machine learning. For machine learning it is useful for creating more effective models with high dimensionality which are less likely to 'overfit' and better generalize with test-set data. It does this by reducing the impact of an effect termed "the curse of dimensionality". Dimension reduction is useful for exploration and visualization of multidimensional data, because humans are limited to effective visualization in higher than 3 dimensions. This technique allows multidimensional data to be "squashed" into lower dimensions, in order for the data to be effectively interpreted.[3] An analogy might be the visualisation of a tree on a television. The television displays a 2D image on a 2D screen, however the tree would be interpreted by the person watching it as a 3D object. Dimensionality reduction in the case of reducing a 23 dimensional data set is not dissimilar. There are several methods of dimensionality reduction, each with their own unique benefits in data exploration.

3.2 An Overview of PCA, LDA and tSNE

PCA or 'principle component analysis' enables identification of strong patterns in complex data sets, and shows clusters of samples based on their similarity. It looks at the data set as a whole and identifies a new set of axis(or components) that capture the most variance in the data. First the data must be standardized, then the eigenvalues and eigenvectors obtained from the covariance matrix by vector decomposition. The eigen values are sorted from largest to smallest resulting in a variance explanation of the data (Scree). Project the components to n number of dimensions, in this case 2 for ease of visualization.PCA looks to maximize the preservation of variance during the dimension reduction process.[4] An example is the reduction of a 2D data set (figure 1a) to a 1D line. If the data points were to be projected directly onto the X axis,

there would be significant loss of the information the variation in the Y axis would provide. A projection onto the Y axis would have the same problem. However a projection onto the principal component 1 (PC1) would optimally preserve information from both X and Y. Principle component 1 (PC1) explains the most variation,and Principle Component 2 (PC2) the second most, as the annotation presents.LDA is similar to PCA as it is also a linear dimension reduction technique, however with the different aim to maximize the separation of the known categories, ie, food groups.LD1 (the first axis that LDA creates) accounts for the most separation between the data, LD2 the second.

t-DistributedStochastic Neighbor Embedding (t-SNE) is a more modern, unsupervised, non-linear technique used predominantly in data exploration and visualizing multi-dimensional data. It functions by preserving small local similarities or "pairwise distances' ', where PCA is concerned with large pairwise distances.[5]

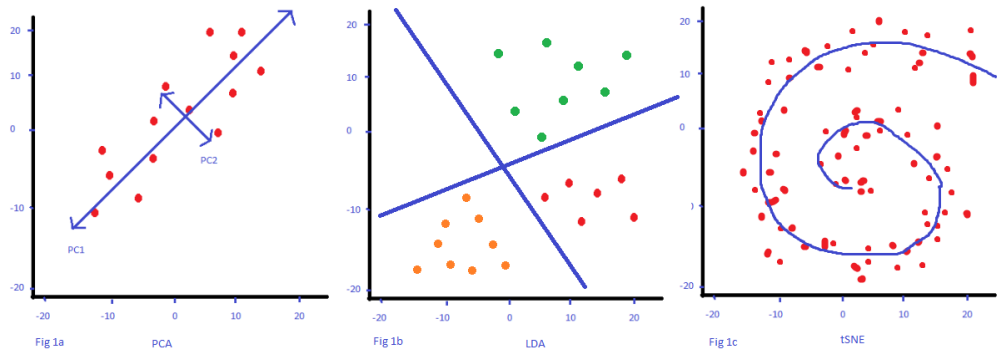

Figure 1: PCA, LDA, tSNE, respectively

3.3 Preprocessing

All preprocessing was implemented with sklearn's preprocessing module, in particular the StandardScaler() function and pandas for dataframe manipulation. It was necessary to amend the data in its raw form before it could be utilized. Firstly, all redundant columns were removed including all USDR columns which were not useful in this study. What remained were 23 columns representing 23 different nutrients for 8618 food samples, where the food groups labels were already provided.

$$Xchanged = \frac{X - \mu}{\sigma} \tag{1}$$

The raw data was not in a state for comparison to be effective. For example the 'Energy-kcal' column was measured in the hundreds, where 'Zinc-mg' was measured between 0-5 mg. For all attributes to be effectively compared,standardisation of the data was required. This allowed for the data to be effectively investigated.

4 Hypothesis 1 - Food groups in the data set would present a unique composition of nutrient ratios per food group

4.1 Original dimensions 2D plots

The first method of data exploration was the implementation of 2D projections of the sample data onto 2 arbitrarily chosen original dimensions (nutrient variables).The data was labelled and color coded by a 'food group.' A sample of random pairs were selected and visualized with a 2D plot. In most cases the 2D plots did not provide much information about the data set. Figure is in an example of this, where, with the exception of 'breakfast cereals', all clusters are overlapping and therefore the plot could not provide much support for the hypothesis. Despite this, it does explain some variance of the clusters, which appear to be explained significantly more so by 'Protein' than 'Riboflavin', as the data is spread in the 'Protein' dimension.In some cases however, valuable information was observed, as shown on figure . Although many of the food groups are still entangled, there is significant clustering and separation of a number of food groups, including 'breakfast cereals' (dark grey), 'sausages and luncheon' (light grey), 'nuts and seeds' (pink), 'spices and herbs' (dark green) and 'snacks' (red). In order to visualize all combinations of original dimension pairs, one would need to analyze a huge number of plots. Not only would this be unfeasible, it would be incomprehensible to visualize the intrinsic complexity of the data as a whole. It is noteworthy that 2D projections on the original dimensions are presenting a biased view of the data, as the trends visualized may be explained by another dimension not present in the 2D plot, despite what it

may suggest. The results of this graph however do support the hypothesis that food groups do have a unique nutrient composition, although not the case for all food groups.

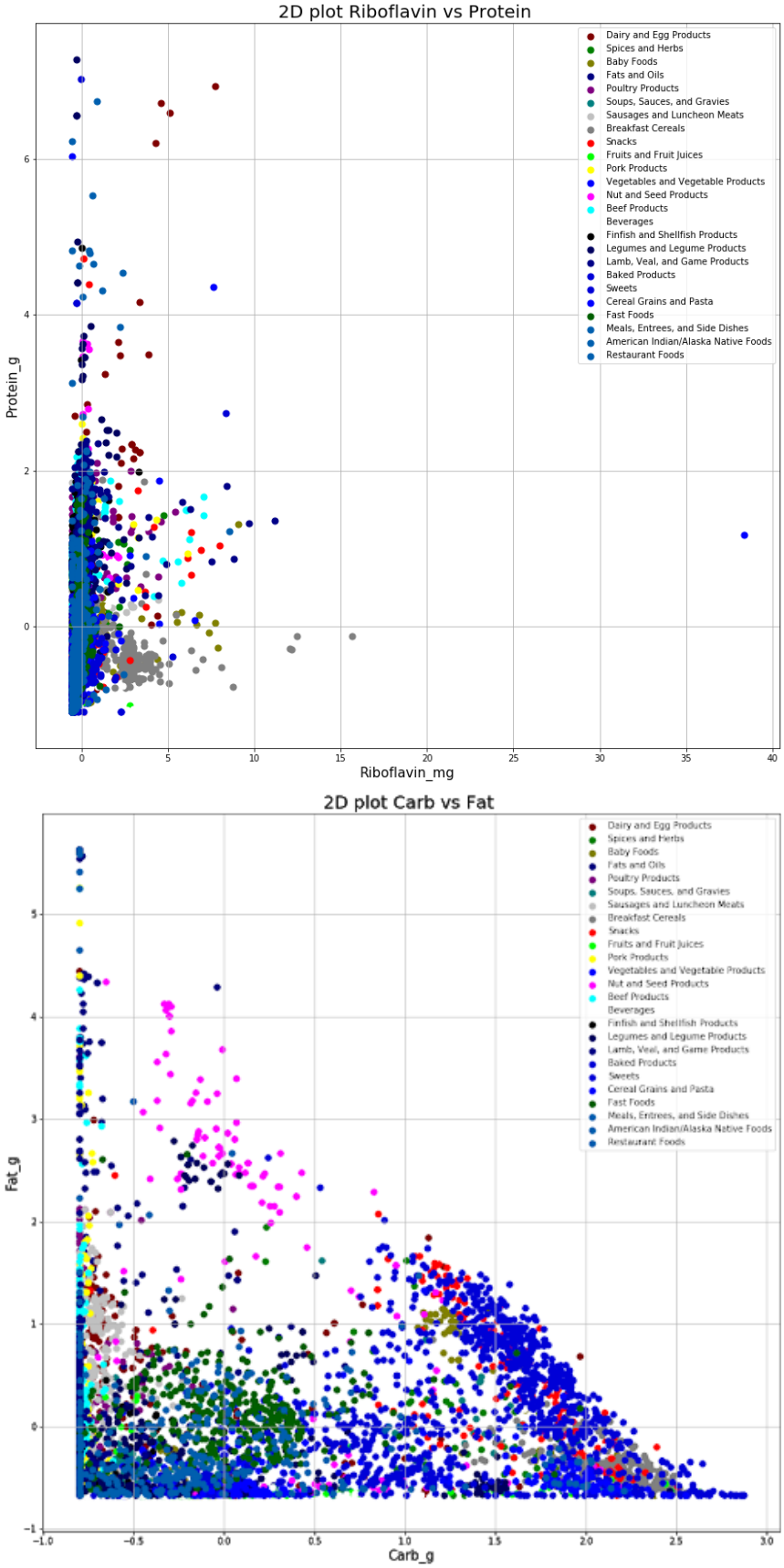

Figure 2: Shows 2D projections original dimensions (2a upper plot, 2b lower plot)

4.2 Principle Component Analysis, PCA

PCA was then implemented to test this hypothesis, as it aims to provide a good overview of the complex data, where trends and relationships may be identified between different food groups.

4.2.1 Variance explanation of principle components

The viability of the principle components were explored, to ensure that the 2 best principle components would adequately explain the variance for projection to reveal significant trends in visualization. The scree plot in figure 3 fig explains this variance of the principle components found in the data set.The first 2 components in the data set explain 37% of the variance, with PC1 = 26%, PC2 = 11%. The top 5 components explain 60% of the variance.

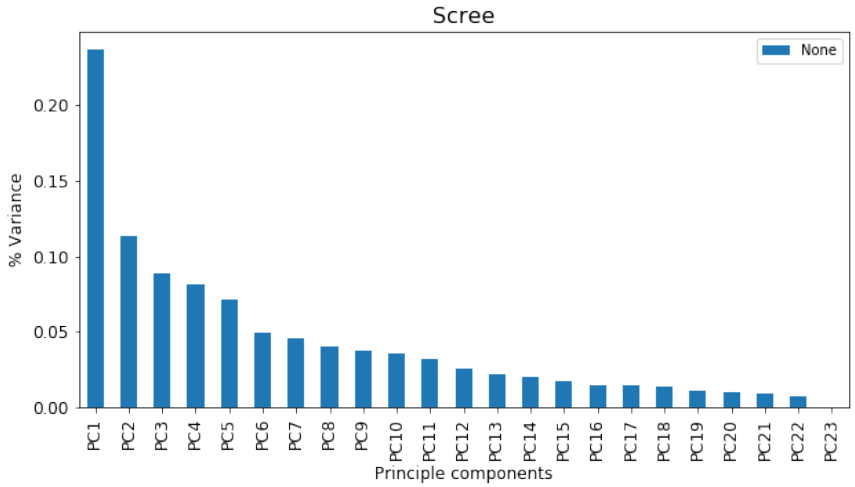

Figure 3: Scree Plot, explanation of variance)

It could be suggested that this is not an ideal outcome, as the projection onto the first two components would only explain 37% of the variance in this case. It is however adequate to potentially identify some trends, and was therefore still explored.

4.2.2 2D PCA plot labelled by food group

The 2D PCA plot presents an entangled and overlapping set of clusters, with the exception of 'breakfast cereals' (grey), which has separated from the main entanglement. PC1 explains the majority of the variance of this cluster, as its elliptical cluster spreads mostly in the PC1 direction. This is indicated with the red elliptical annotation on the figure. Its variance is described less in PC2. PC2 predominantly explains the variation of the large blue cluster(sweets, grains, pasta and baked products) and the turquoise and green clusters (meals) better than PC1. This is because the elliptical spread is in the direction of PCA 2. The orange annotation on the graph indicates this. Although the majority of food groups were not separated by PCA with PC1 and PC2, it did however enable separation of some clusters. The projection of all clusters onto PC1 would result in good separation between grey and both blue, turquoise and green clusters.It is mostly likely a separation of 'grey' and the other clusters hidden in the entanglement, but remains unseen in this plot. Furthermore, PC1 would not separate 'nuts and seeds' (pink) and 'breakfast cereals' as this projection would result in significant overlap.If the data were to be projected onto PC2 however, there would be clear separation of both 'nuts and seeds' and 'breakfast cereals'. It is noteworthy that PCAs purpose is not necessarily to separate clusters within data, but to identify strong patterns in complex data sets. However this can be a useful outcome, and often a subsequent result of projections on principle components which aim to maximize variance preservation.

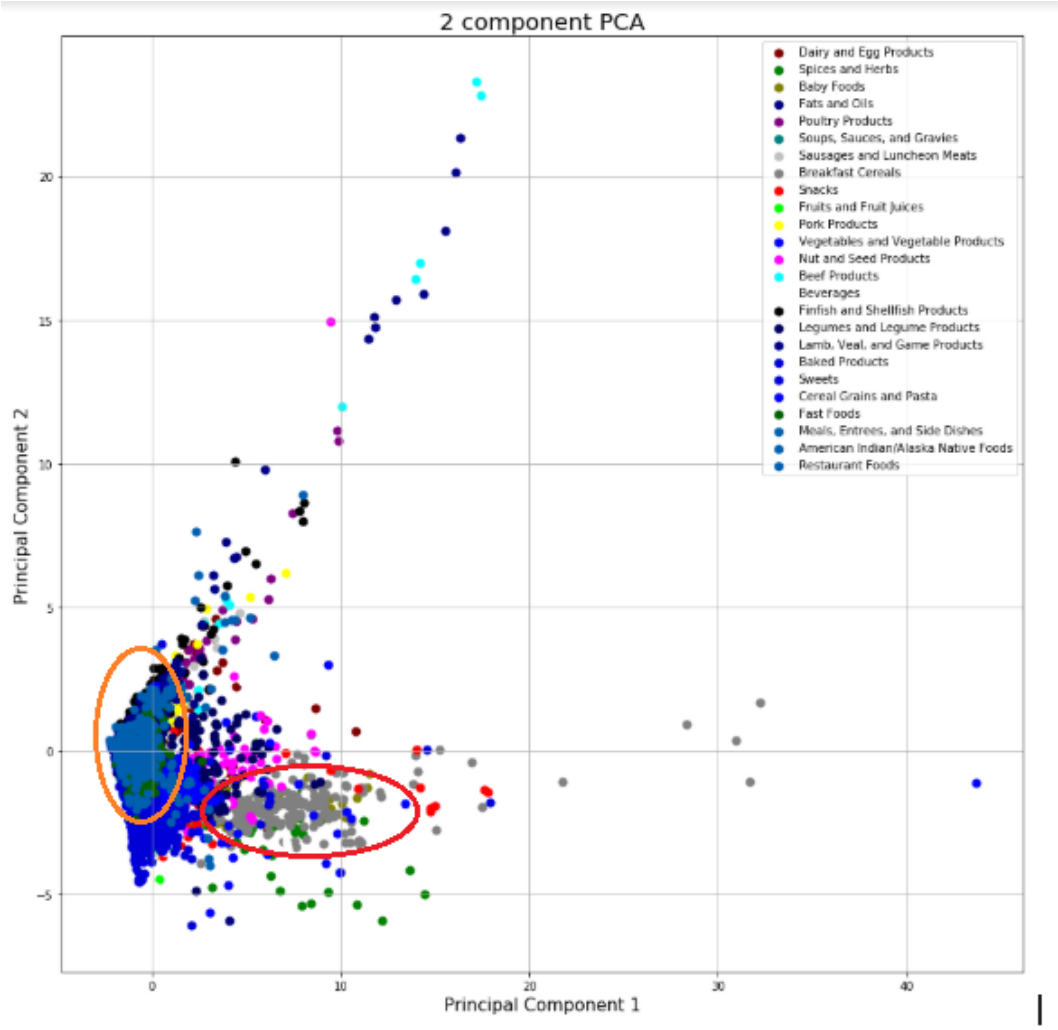

Figure 4: 2D PCA, PC1 and PC2, Food group labelled)

The performance of the 2D PCA plot, although did show clustering and trends, was overall unsuccessful in proving the stated hypothesis, despite clustering, some variance explanation and some separation of clusters. This was due to the majority of the data overlapping, and therefore impossible to identify unique trends of each food group. That being only the first two PCs were plotted, where perhaps other components could be explored. It may be suggested that the unsatisfactory outcome of the PCA plot may be due to the low percentage variance explanation for PC1 and PC2 of only 37%. In order to better support the hypothesis, separated elliptical clusters need to have been present, which is why other methods were necessary to explore the data further.

4.3 LDA

Due to PCA presenting the issue of overlapping classes, it was reasoned a better approach may be to implement LDA, in an attempt visualize the data as distinct separate classes. This is because its aim is to perform dimension reduction with maximal separation of the classes, in which may better support the hypothesis. Once again however, the LDA plot presents similarities with the PCA plot where the majority of the food group clusters overlap, and was therefore unsuccessful. It did not present much more information that PCA did not already provide. Despite this, it is noteworthy that like the PCA plot, breakfast cereals are once again separated successfully. Furthermore, LD2 successfully separated breakfast cereals from all other food groups. LD1 successfully separated 'baked products' and 'fast food' from breakfast cereals.For both PCA and LDA it may be suggested that they are limited as they are both linear methods of dimension reduction, and may also explain why they are producing similar results. The distinct separation of breakfast cereals with both methods may be of some interest to a nutritionist. This reinforces a notion that breakfast cereals have an especially unique composite of nutrients in comparison to the other food groups. It should be noted, that due to the large number of classes, it was difficult to visualize, particularly with many colors (and similar colors).

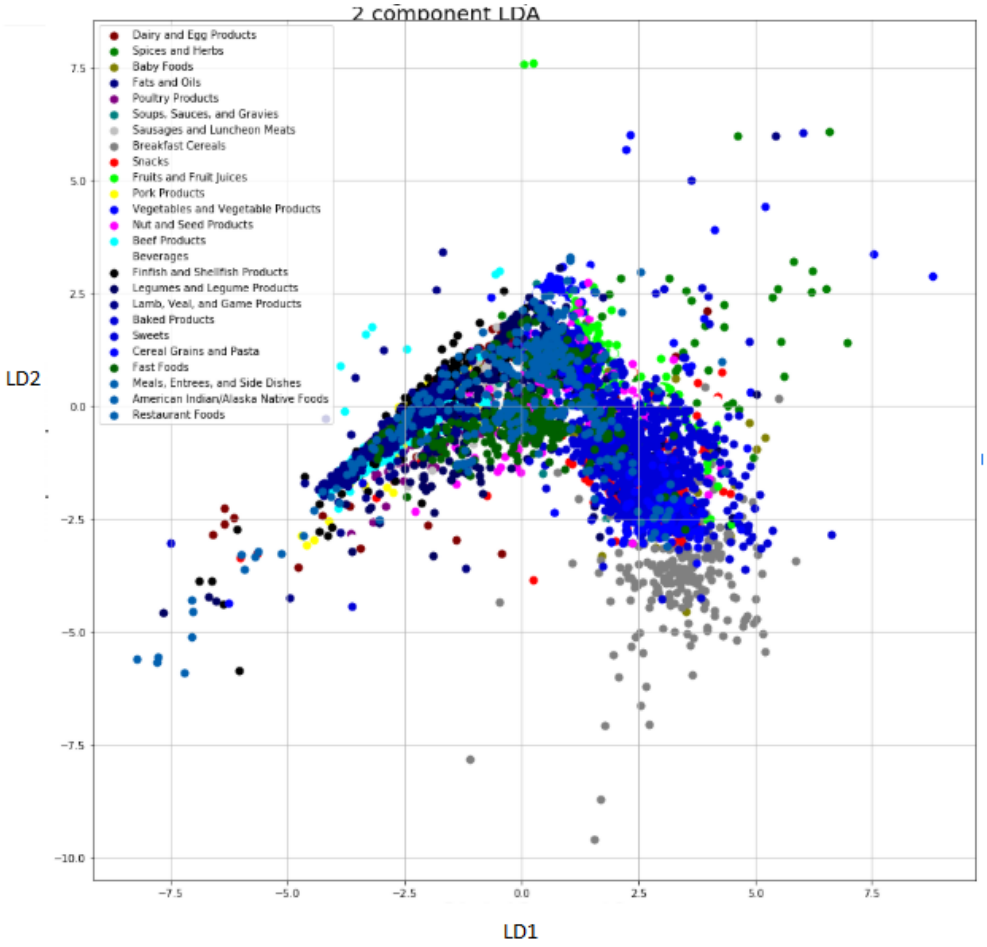

Figure 5: 2D LDA, PC1 and PC2, Food group labelled)

4.3.1 tSNE

Due to both LDA and PCA being unable to separate the clusters successfully, a nonlinear and more modern method of dimension reduction was implemented. The results were significantly more successful, as all food groups had distinct clusters on the most part, and were separated as shown in figure. Despite there being some overlap, and some clusters being slightly dispersed, it allowed the food groups to be separated by their unique composition of properties for the most part, and therefore supports the hypothesis.The hyper-parameter values selected for 'Perplexity' was set to 30 and 'Learning Rate' set to 150, which showed the best clustering and separation from the other parameters tested, through trial and error.

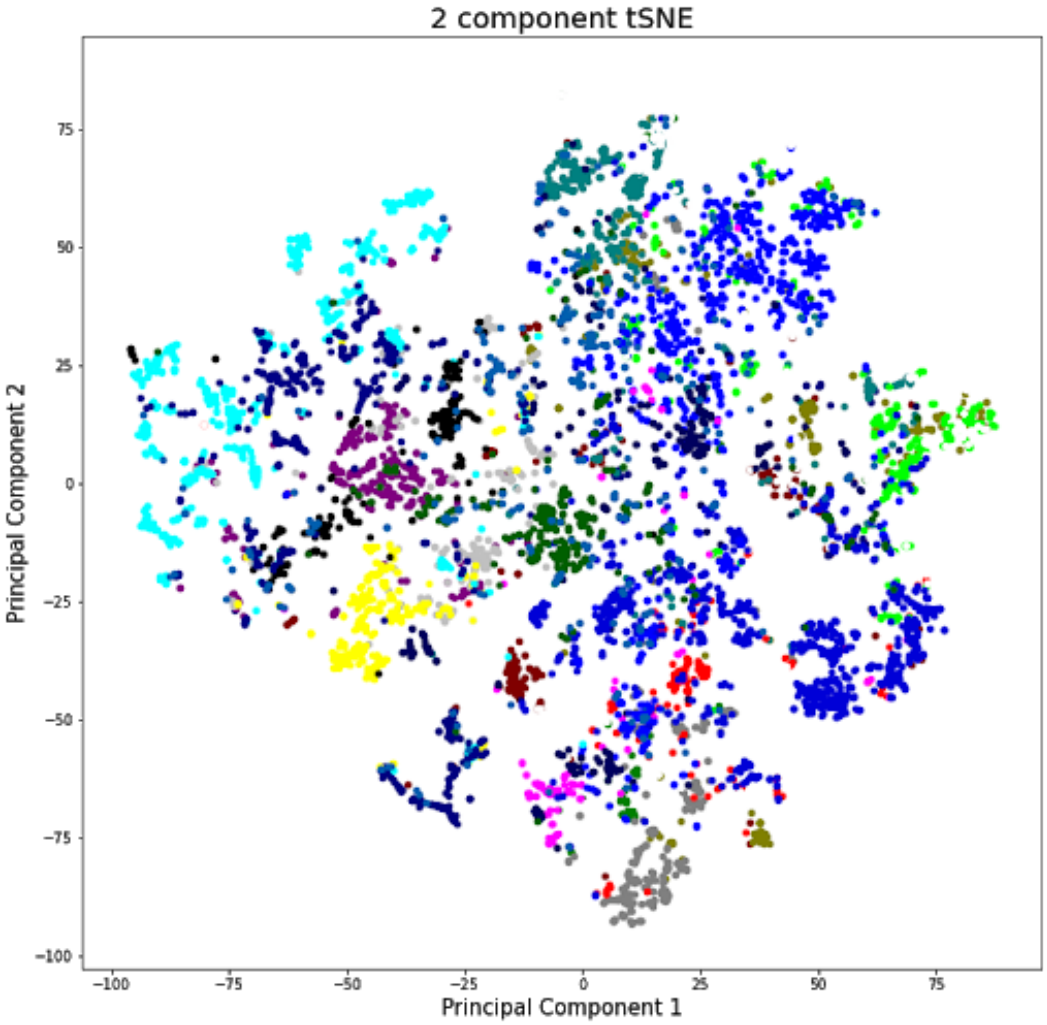

Figure 6: tSNE, Food group labelled)

5 Hypothesis 2 - Foods with high fat content and high carbohydrate content would be indicative of high levels of energy

Principal components are a combination of varied influences of the original dimensions (in this case nutrient measurements of food samples).

These influences can be visually explained in a loading plot, which presents how each original dimension influences each of the principal components that the data is projected on, and the differences among the clusters in the data set.

A bi-plot is a combination of a loading plot overlaid onto a PCA plot.

The direction of a dimension in a loading plot is indicated by the direction on the lines. It portrays which PC it primarily influences, where the length of the line, indicates how strongly it influences the principle component in that direction. When two vectors are close together and forming a more acute angle, the variables (in this case nutrients) are positively correlated with each other. When two vectors are 90 degrees to each other, the variables are unlikely to be correlated. When two vectors tend towards 180 degrees from each other, they are likely to be negatively correlated.[6]

Magnesium and carbohydrates as indicated in the pink circle annotation in figure are likely to be positively correlated to one another, and explain the majority of the variance in PC2. Fat and energy are positively correlated with one another and also significantly explain the variance in PC2 also, however in the opposite direction (purple circle). This suggests Energy and Fat, are negatively correlated with Magnesium and carbohydrates. Both would not likely have much correlation with protein, riboflavin and iron (orange circle) and vitamin b12 and selenium (green circle), as they both tend to 90 degree from them.

All nutrients in both orange and green circles do not explain the variance in PC2, in particular Iron which is almost exactly 90 degrees to PC2. However they do explain much of the variance in PC1 as they tend to that direction. It's noteworthy that there is no indication of any negatively correlating variable in PC1.

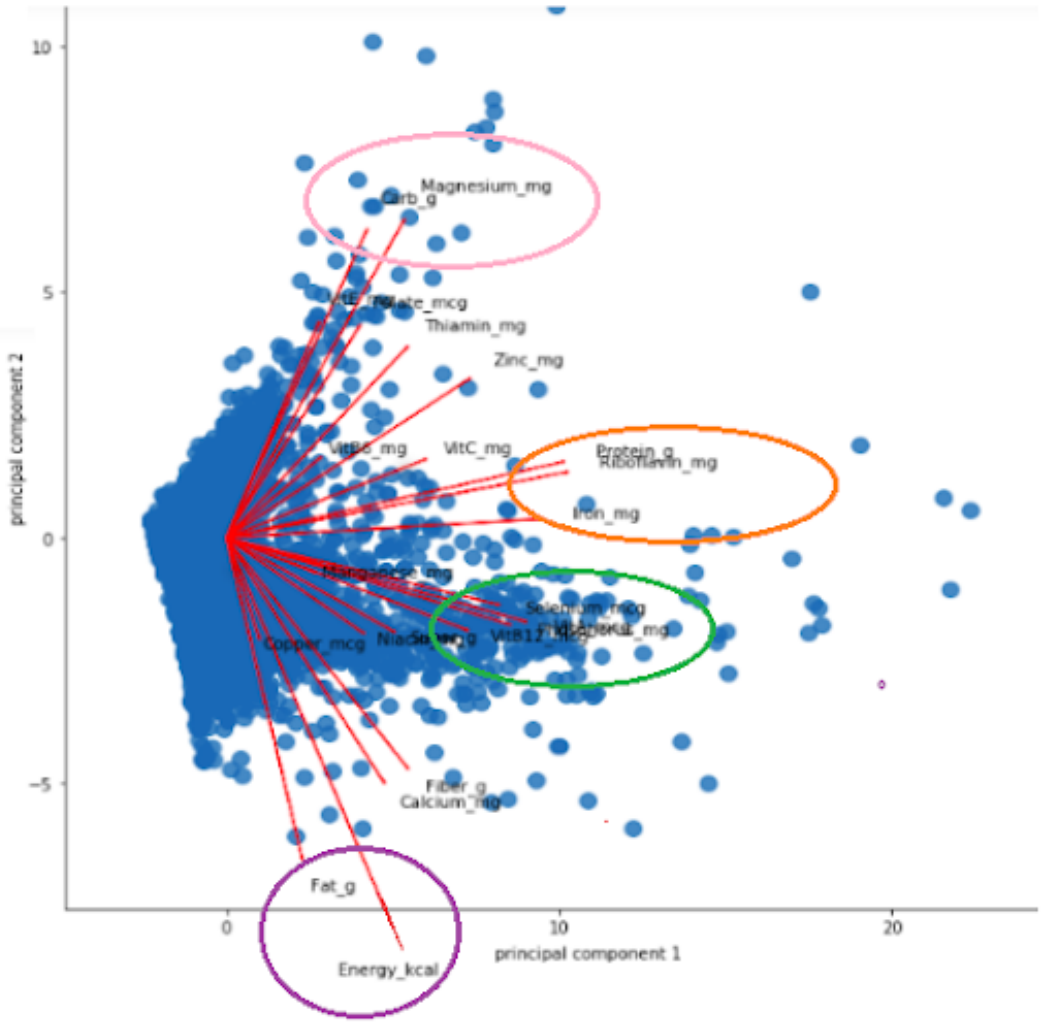

Figure 7: Bi-plot PC1, PC2

What is expected is the positive correlation Fat-g and energy, with fat being known to be energy rich. What is not expected is that energy and carbs are negatively correlated. Although fats do have twice as much calories that carbs per gram. Despite this, it would be expected that fat and carbohydrates would be expected to be highly negatively correlated, and perhaps this is what the plot is expressing more strongly with PC1 and PC2. This perhaps shows the limitations of loading plots. Due to this, the eigenvalues and eigenvectors were visualized in 1D on a bar chart, in order to provide a deeper understanding of the data and underlying trends.

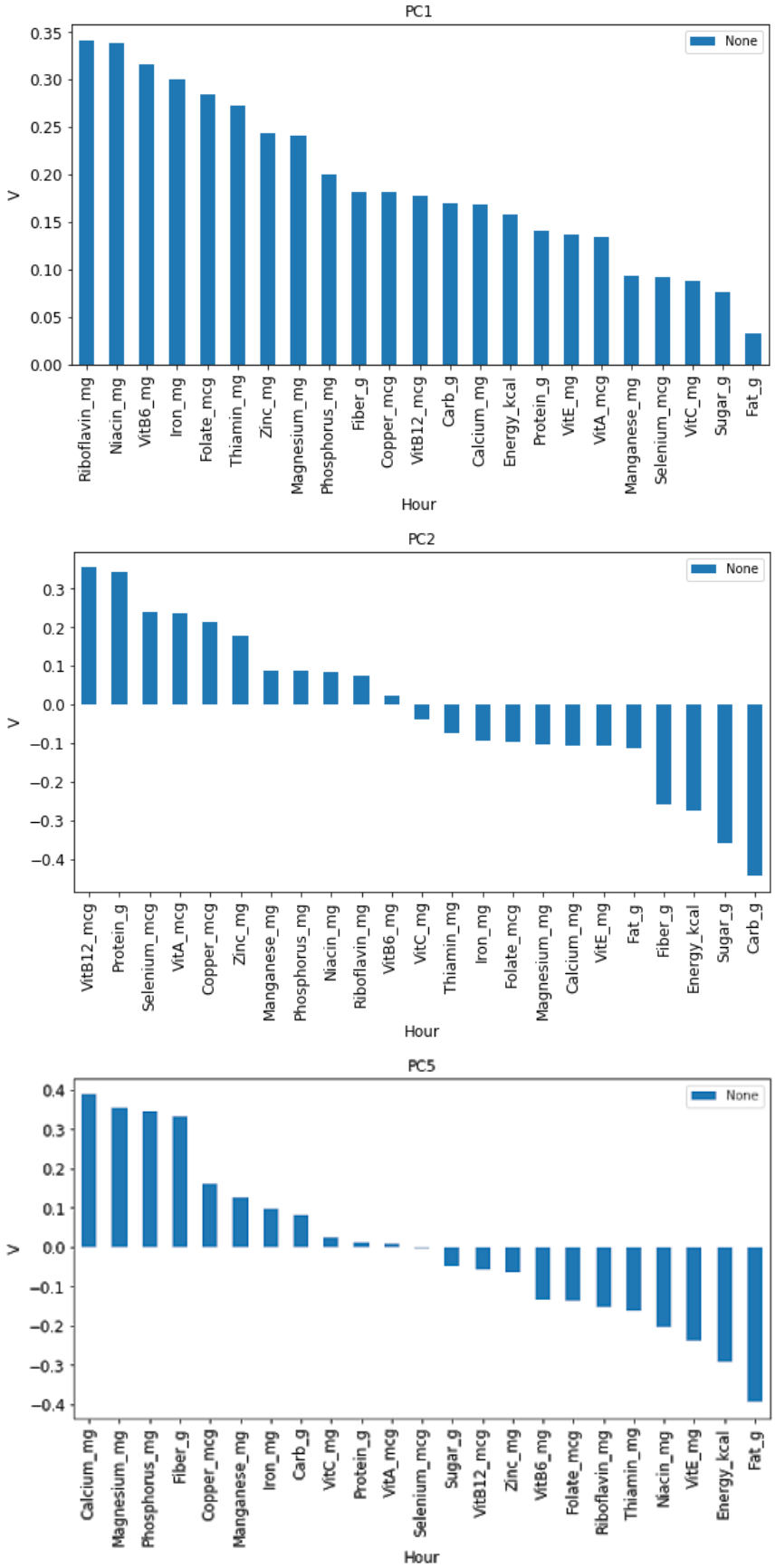

Figure 8: PC1, PC2, PC3

Like the plot of PC1 vs PC2, PC3 also presents 'Fats' and 'Energy' being strongly correlated. P3 however, unlike the

plot, presents strong correlation of 'Energy' and 'carbohydrates', therefore supporting the hypothesis.[7]

Riboflavin and protein are positively correlated, which also would be expected, as riboflavin itself is a protein, is involved in the breakdown of other proteins, and is found in high quantities in high protein rich foods, such as organ meats, egg whites, adn lean meats.[8]

Overall, there is strong evidence supporting the hypothesis, however the complex dynamic of all dimensions, requires more than the first 2 principle components to be explored. This is more the case, as there is only 37% of the variance explained in the first 2 components.

5.1 Hypothesis 3 - It would be possible to correctly predict Energy (kcal) within an adequate range, exclusively from the other 22 of the 23 nutrients (dimensions) into the class of low, medium, high, and very high energy levels

5.1.1 Preprocessing

From the already preprocessed and standardised data set as previously described, the data was further processed further in order to investigate this hypothesis. The 'Energy-kcal' attribute was first separated from the other 22 dimensions as set Y. The remaining variables were held as variable X. The 'food group' labels were then removed from the data set, to allow for the labeling scheme 'Energy-level'. The classes and labelling scheme was created by first using the data set Y ('Energy Levels') to classify the samples into 'low'(blue), 'med'(green), 'high'() yellow and 'very-high' (red). The range these classes represented were reasoned arbitrarily from the histogram in fig. This concluded the further preprocessing necessary for testing the hypothesis. PCA, LDA and tSNE were once again implemented on the data set X. The results are shown in figure.

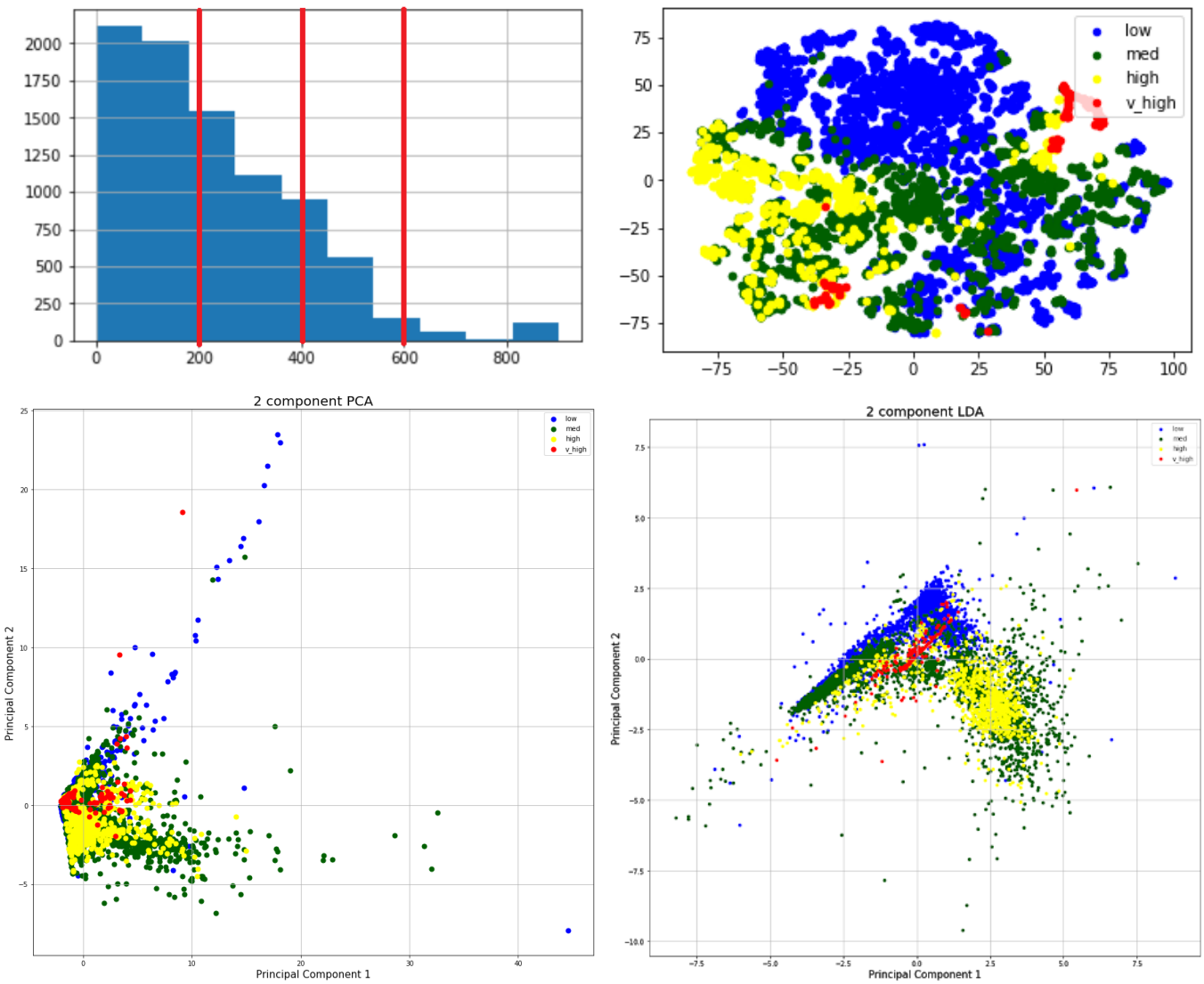

Figure 9: PC1, PC2, PC3

The removal of the energy attribute did not alter the data structure in a significant manner, as looked visually similar, as the the previous results with all 23 dimensions. Despite distinct clustering, LDA and PCA were once again unable to separate them. PCA was particularly unsuccessful, with separation not possible in both PC1 and PC2, although it did explain much of the variance of 'medium' energy level class, in the PC1, and low energy level in PC2 ,as clusters data were spread in these directions, respectively. LDA performed slightly better with separation possible of 'low vs high' and 'high vs very high' in LD1and LD2, however was not possible with any separation due to significant overlap. tSNE, although showed the best separation, however there was still a degree of overlap, with 'high and low', 'very high appeared too dispersed for it to be used to make just conclusions, forming two small clusters. ' It was however able to separate yellow and blue, and yellow and red.

Inclusion, at it was not possible to categorically accept the hypothesis, due to the lack of distinct and separated classes in any of the techniques implemented.

5.2 Conclusion

In conclusion, hypothesis 1 was supported, as the results support the hypothesis that food groups do have a unique nutrient composition, although this is not the case for all food groups. There was strong evidence in support of hypothesis 2, when all tests are cumulatively taken into consideration. Hypothesis 3 was not sufficiently supported as it was not possible to obtain distinct separation of clusters with all dimension reduction techniques, which was required to adequately support this hypothesis.

References

[1] U.S. DEPARTMENT OF AGRICULTURE. *USDA*. Agricultural Research Service, United Nations, accessed 1 March 2020, https://fdc.nal.usda.gov/

[2] Craig Kelly, USDA database, accessed 1 March2020, https://data.world/craigkelly/usda-national-nutrient-db

[3] Nguyen LH, Holmes S (2019) Ten quick tips for effective dimensionality reduction. PLoS Comput Biol 15(6): e1006907. https://doi.org/10.1371/journal.pcbi.1006907

[4] Jonathon Shlens, A Tutorial on Principal Component Analysis, Google Research Mountain View, CA 94043 (Dated: April 7, 2014; Version 3.02)

[5] Laurens van der Maaten, Visualizing Data using t-SNE, Geoffrey Hinton; 9(Nov):2579–2605, 2008

[6] Greenacre, M. Principal component analysis biplots. in Biplots in Practice 59–67 (Fundación BBVA, 2010).

[7] Cleveland Clinic, Fat and Calories, date accessed 2 march 2020, https://my.clevelandclinic.org/health/articles/4182-fat-and-calories

[8] U.S. Department of Health Human Services, National Institutes of Health, date accessed 2 march 2020, https://ods.od.nih.gov/factsheets/Riboflavin-HealthProfessional/