Principal Component Analysis Applied to USDA National Nutrient Data Uncovering Multivariate Patterns in Nutritional Composition Across Different Food Products

Abstract

This project employs Principal Component Analysis (PCA) to examine the correlation between the nutritional content of various food products derived from the USDA nutritional dataset. Specifically, it involves randomly selecting pairs of vitamins and minerals from each output, interpreting the meaning of the correlation between them, and demonstrating how this meaning varies with different types of correlation. PCA simplifies nutritional data analysis by reducing its dimensionality while preserving most of the original variability. The significance of this research is to provide dieticians and public health advisors with statistical evidence to influence the public on how and why to eat healthier.

Introduction

There has been various literature on the topic of analyzing the nutritional content of food by utilizing Principal Component Analysis. The International Journal of Food and Nutrition had similarly applied PCA to reveal food patterns by comparing the efficacy of scree plot visual inspection. Their research, “Food or nutrient pattern assessment using the principal component analysis applied to food questionnaires”, observed clear discontinuity after the third eigenvalue when simulating three food patterns, which indicated scree plot visual inspection to be more suitable for food pattern identification (Panagiotakos et al., 2007). This research provides evidence of PCA use in the field of nutrition.

On a more grand scale, studying nutrition habits and diseases could hold significant practical value by demonstrating the relationship between disease and the correlation of patients' food habits and patterns. A study published in the Journal of the American Dietian Association analyzed how dietary patterns have been associated with metabolic syndrome. The study ran PCA on a cross-sectional survey. Six components were identified, which explained 56% of the intake variation. The study explained how each food category had either a negative or positive association with metabolic syndrome (Ricci et al., 2019). Against this backdrop and evidence of PCA's wide use in health and nutrition, this research aims to explore the correlation between the nutritional profiles of various food categories. By leveraging the USDA's National Nutrient Database, we delve into the relationships between essential nutrients found in commonly consumed foods. The focus is to also distinguish the underlying patterns and relationships uniform among key vitamins and minerals across different food categories.

Methodology

The data set used in this project was obtained from the USDA's National Nutrient Database, a comprehensive resource providing detailed information on the nutritional content of numerous US food items. The dataset encompasses information on 25 variables where 23 different nutrients across 8,618 observations. From these, 16 key nutrients were selected for analysis using a method known as principal component analysis, including Vitamins A, B6, B12, C, E, Folate, Niacin, Riboflavin, Thiamin, Calcium, Copper, Iron, Magnesium, Phosphorus, Selenium, and Zinc. These variables were specifically chosen due to their essential role in human health and their prevalence in various foods, aiming to uncover insights into the nutritional profiles of different food groups. Within the dataset, foods are categorized into groups such as egg & dairy, vegetables, meats, etc on which the analysis is primarily focused. The selected food categories - egg & dairy, poultry, and sweets were chosen because of their common consumption in the typical American diet, and their diverse dietary components. After analyzing the appropriate number of eigenvalues to select and explain how successful it is in explaining the variances, our research will provide some evidence into scientific reasoning of why various nutrients have teamwork relationships in the human body. Through analyzing these categories, the aim is to demonstrate the intricate relationship between the nutritional composition, thereby providing insights to inform dietary recommendations.

Finding

In analyzing diary and eggs, the initial three eigenvalues account for approximately 84% of the variability. The primary principal component demonstrates a positive relationship with 13 key variables, such as Vitamins A, B6, B12, C, Folate, Niacin, Riboflavin, Thiamin, Copper, Iron, Phosphorus, Selenium, and Zinc, indicating their simultaneous increase. Similarly, the second principal component shows a positive correlation with Vitamin C and Copper, while the third principal component is positively associated with Vitamin E and Thiamin.

For poultry, roughly 90% of the variance is explained by the top three eigenvalues. The leading principal component exhibits a positive correlation with 13 variables, mirroring those in the previous category. Additionally, the second principal component demonstrates a slight positive correlation with Vitamin C and Copper, while the third principal component shows a positive link with Vitamin E and Thiamin.

Regarding sweets, approximately 70.12% of the variation is accounted for by the first three eigenvalues. The foremost principal component shows a positive correlation with 7 original variables, including Vitamin E, Folate, Niacin, Thiamin, Copper, Iron, Magnesium, Phosphorus, and Zinc. Meanwhile, the second principal component displays a negative correlation with Vitamin A, alongside positive correlations with Vitamin C and Calcium. Lastly, the third principal component is positively correlated with Vitamin A, Vitamin B12, Riboflavin, and Calcium.

Discussion

The results underscore the crucial role of eggs and dairy in supplying essential nutrients such as Vitamin A and zinc, vital for overall health. According to the National Library of Medicine, Vitamin A supports vision and mucosal barrier function, while zinc bolsters immunity and facilitates the transportation of Vitamin A. The concurrent presence of these nutrients in eggs and dairy highlights their nutritional significance.

Similarly, poultry products provide a unique blend of Vitamin E and Selenium. As stated by the National Library of Medicine, Vitamin E shields cells from oxidative damage, while Selenium augments antioxidant capacity. This synergy in poultry helps combat oxidative stress and promotes cellular health.

Furthermore, research indicates that Vitamin A absorption is optimized when consumed alongside fats, as suggested by the National Health Service. Conversely, the consumption of sweets, particularly low-fat options, may impede Vitamin A absorption.

These findings deepen our comprehension of the nutritional value of these food groups and underscore the importance of dietary synergy in achieving optimal health outcomes.

Conclusion

The study has revealed an important connection between various vitamins and minerals, indicating a collaborative relationship. Investigating this partnership can significantly benefit dietitians in offering informed advice on nutrition. Principal Component Analysis (PCA) can help uncover patterns in the nutritional composition of food items. Also, certain principal components are primarily influenced by mineral levels. The strong correlation between different vitamins and minerals suggests their synergistic action within the body. For instance, Vitamin E and selenium function together as a potent duo against harmful oxidation. Exploring these cooperative relationships could enhance dietary guidance. Furthermore, negative correlations may indicate a balance mechanism where an increase in one vitamin corresponds to a decrease in another. Understanding these correlations provides insight into how the body regulates nutrient levels to maintain equilibrium. Besides, studying these teamwork relationships could lead to more precise dietary recommendations, helping individuals make healthier choices.

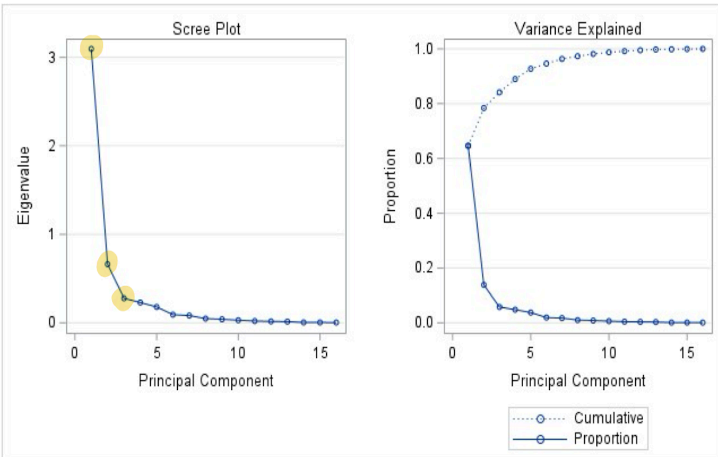
Appendix

```
proc import out=Nutrient datafile="C:\Users\bushr\OneDrive\Desktop\nndb_flat.csv"
dbms=csv replace; getnames=yes;
*PCA;
data Nutrient;
    set Nutrient;
VitA_mcg = log10(VitA_mcg);
VitB6_mg = log10(VitB6_mg);
VitB12_mcg = log10(VitB12_mcg);
VitC_mg = log10(VitC_mg);
VitE_mg = log10(VitE_mg);
Folate_mcg = log10(Folate_mcg);
Niacin_mg = log10(Niacin_mg);
Riboflavin_mg = log10(Riboflavin_mg);
Thiamin_mg = log10(Thiamin_mg);
Calcium_mg = log10(Calcium_mg);
Copper_mcg = log10(Copper_mcg);
Iron_mg = log10(Iron_mg);
Magnesium_mg = log10(Magnesium_mg);
Phosphorus_mg = log10(Phosphorus_mg);
Selenium_mcg = log10(Selenium_mcg);
Zinc_mg = log10(Zinc_mg);
run;
proc sort data=Nutrient;
  by FoodGroup;
where FoodGroup in ('Dairy and Egg Products', 'Poultry Products', 'Sweets');
run;
title 'PCA on USDA National Nutrient Data';
proc princomp data=Nutrient cov out=n;
by  FoodGroup;
where FoodGroup in ('Dairy and Egg Products', 'Poultry Products', 'Sweets');
var VitA_mcg VitB6_mg VitB12_mcg VitC_mg VitE_mg Folate_mcg Niacin_mg Riboflavin_mg Thiamin_mg Calcium_mg
Copper_mcg  Iron_mg Magnesium_mg Phosphorus_mg Selenium_mcg Zinc_mg;
run;

proc corr data=n noprob;
by  FoodGroup;
where FoodGroup in ('Dairy and Egg Products', 'Poultry Products', 'Sweets');
var prin1 prin2 prin3 VitA_mcg VitB6_mg VitB12_mcg VitC_mg VitE_mg Folate_mcg Niacin_mg Riboflavin_mg Thiamin_mg Calcium_mg
Copper_mcg  Iron_mg Magnesium_mg Phosphorus_mg Selenium_mcg Zinc_mg;
run;
```

Principal Component SAS Output for Dairy and Eggs

Total Variance	4.7962597058

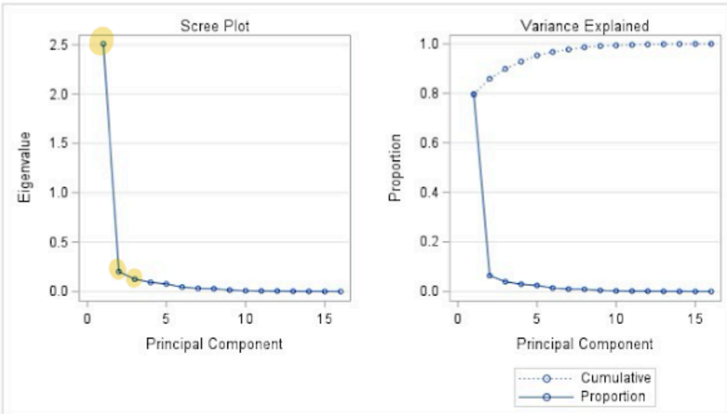


Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	3.09599788	2.43347239	0.6455	0.6455
2	0.66252549	0.38521277	0.1381	0.7836
3	0.27731272	0.04757075	0.0578	0.8415
4	0.22974197	0.04978344	0.0479	0.8894
5	0.17995853	0.08788103	0.0375	0.9269
6	0.09207749	0.01003059	0.0192	0.9461
7	0.08204690	0.03445696	0.0171	0.9632
8	0.04758994	0.00805200	0.0099	0.9731
9	0.03953794	0.01028141	0.0082	0.9813
10	0.02925653	0.00912516	0.0061	0.9874
11	0.02013137	0.00418270	0.0042	0.9916
12	0.01594867	0.00268841	0.0033	0.9950
13	0.01326027	0.00933487	0.0028	0.9977
14	0.00392540	0.00008335	0.0008	0.9986
15	0.00384205	0.00073551	0.0008	0.9994
16	0.00310654		0.0006	1.0000

	Prin1	Prin2	Prin3
Prin1	1.00000 64	0.00000 64	0.00000 64
Prin2	0.00000 64	1.00000 64	0.00000 64
Prin3	0.00000 64	0.00000 64	1.00000 64
VitA_mcg	0.57547 64	-0.55031 64	0.57563 64
VitB6_mg	0.92928 64	0.17363 64	-0.03825 64
VitB12_mcg	0.79970 64	0.41643 64	0.13926 64
VitC_mg	0.78092 64	-0.15734 64	-0.24785 64
VitE_mg	0.57507 64	-0.72067 64	-0.19380 64
Folate_mcg	0.90784 64	0.04657 64	-0.16191 64
Niacin_mg	0.95346 64	0.04481 64	-0.11804 64
Riboflavin_mg	0.70661 64	0.44207 64	0.27209 64
Thiamin_mg	0.90896 64	0.16920 64	-0.04367 64
Calcium_mg	0.72786 64	0.42719 64	0.32335 64
Copper_mcg	0.81315 64	0.02756 64	-0.15029 64
Iron_mg	0.89688 64	-0.01919 64	-0.02451 64
Magnesium_mg	0.91276 64	0.25463 64	0.05391 64
Phosphorus_mg	0.76442 64	0.40261 64	0.27592 64
Selenium_mcg	0.77709 64	0.37165 64	0.02808 64
Zinc_mg	0.92000 64	0.24742 64	0.08200 64

Principal Components SAS Output for Poultry



Total Variance	3.1539945994

Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	2.50970865	2.30834742	0.7957	0.7957
2	0.20136123	0.07588615	0.0638	0.8596
3	0.12547508	0.03222970	0.0398	0.8993
4	0.09324539	0.01631332	0.0296	0.9289
5	0.07693207	0.03230416	0.0244	0.9533
6	0.04462791	0.01316678	0.0141	0.9675
7	0.03146113	0.00273902	0.0100	0.9774
8	0.02872211	0.01353785	0.0091	0.9865
9	0.01518425	0.00689274	0.0048	0.9914
10	0.00829151	0.00194484	0.0026	0.9940
11	0.00634667	0.00103247	0.0020	0.9960
12	0.00531420	0.00209055	0.0017	0.9977
13	0.00322365	0.00094255	0.0010	0.9987
14	0.00228110	0.00108408	0.0007	0.9994
15	0.00119702	0.00057440	0.0004	0.9998
16	0.00062262		0.0002	1.0000

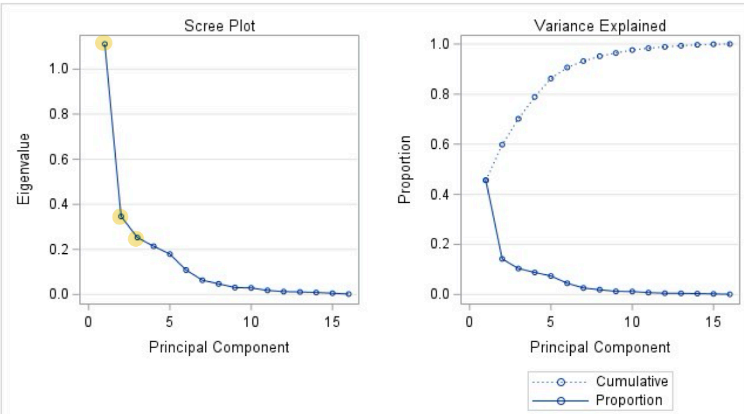
	Prin1	Prin2	Prin3
Prin1	1.00000 32	0.00000 32	0.00000 32
Prin2	0.00000 32	1.00000 32	0.00000 32
Prin3	0.00000 32	0.00000 32	1.00000 32
VitA_mcg	0.95101 32	-0.26763 32	0.05386 32
VitB6_mg	0.58573 32	-0.13081 32	0.38241 32
VitB12_mcg	0.94109 32	-0.00566 32	-0.28695 32
VitC_mg	0.77813 32	0.46307 32	-0.09099 32
VitE_mg	0.32521 32	-0.23610 32	0.41915 32
Folate_mcg	0.97570 32	-0.08812 32	0.02910 32
Niacin_mg	0.64291 32	-0.13070 32	0.12085 32
Riboflavin_mg	0.96367 32	0.08102 32	0.03481 32
Thiamin_mg	0.62569 32	0.26921 32	0.65145 32
Calcium_mg	0.07097 32	0.11791 32	-0.24919 32
Copper_mcg	0.82721 32	0.44250 32	0.16582 32
Iron_mg	0.81397 32	0.32008 32	0.05197 32
Magnesium_mg	0.23353 32	0.08215 32	0.23829 32
Phosphorus_mg	0.76024 32	0.19092 32	0.26780 32
Selenium_mcg	0.88560 32	0.02584 32	-0.26004 32
Zinc_mg	0.74840 32	0.35338 32	-0.20510 32

Principal Components SAS Output for Sweets

Total Variance	2.437824843

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	1.11106419	0.76498371	0.4558	0.4558
2	0.34608047	0.09372054	0.1420	0.5977
3	0.25235993	0.03866013	0.1035	0.7012
4	0.21369980	0.03469778	0.0877	0.7889
5	0.17900202	0.07100695	0.0734	0.8623
6	0.10799507	0.04515431	0.0443	0.9066
7	0.06284076	0.01632614	0.0258	0.9324
8	0.04651462	0.01590690	0.0191	0.9515
9	0.03060771	0.00170146	0.0126	0.9640
10	0.02890625	0.01075186	0.0119	0.9759
11	0.01815439	0.00542580	0.0074	0.9833
12	0.01272859	0.00138454	0.0052	0.9886
13	0.01134405	0.00253161	0.0047	0.9932
14	0.00881245	0.00306952	0.0036	0.9968
15	0.00574292	0.00377129	0.0024	0.9992
16	0.00197163		0.0008	1.0000



	Prin1	Prin2	Prin3
Prin1	1.00000 56	0.00000 56	0.00000 56
Prin2	0.00000 56	1.00000 56	0.00000 56
Prin3	0.00000 56	0.00000 56	1.00000 56
VitA_mcg	-0.32530 56	-0.69713 56	0.58151 56
VitB6_mg	0.44050 56	0.38964 56	0.45540 56
VitB12_mcg	-0.02623 56	0.46204 56	0.65245 56
VitC_mg	-0.21040 56	0.52892 56	0.12106 56
VitE_mg	0.75686 56	-0.01998 56	-0.15722 56
Folate_mcg	0.73605 56	0.21297 56	0.29291 56
Niacin_mg	0.87260 56	0.22124 56	0.16748 56
Riboflavin_mg	0.02245 56	0.38367 56	0.60419 56
Thiamin_mg	0.73365 56	0.33151 56	0.29679 56
Calcium_mg	0.05012 56	0.54947 56	0.53491 56
Copper_mcg	0.83867 56	-0.34595 56	-0.07116 56
Iron_mg	0.79406 56	-0.32741 56	0.01569 56
Magnesium_mg	0.91513 56	-0.17190 56	0.00151 56
Phosphorus_mg	0.68119 56	0.26837 56	0.47664 56
Selenium_mcg	0.12281 56	0.07156 56	0.25777 56
Zinc_mg	0.82230 56	0.04734 56	0.16610 56

References

1. Panagiotakos , Demosthenes B., et al. “The Association between Food Patterns and the Metabolic Syndrome Using Principal Components Analysis: The Attica Study.” *Journal of the American Dietetic Association*, Elsevier, 23 May 2007, www.sciencedirect.com/science/article/pii/S0002822307004336.

2. “Principal Components Exercise1 - Dataset by Exercises.” *Data.World*, 19 Sept. 2017, data.world/exercises/principal-components-exercise-1.

3. Ricci, Cristian, et al. “Food or nutrient pattern assessment using the principal component analysis applied to food questionnaires. pitfalls, tips and tricks.” *International Journal of Food Sciences and Nutrition*, vol. 70, no. 6, 22 Feb. 2019, pp. 738–748, https://doi.org/10.1080/09637486.2019.1566445.

4. Rahman, Mohammad M, et al. “Synergistic Effect of Zinc and Vitamin a on the Biochemical Indexes of Vitamin a Nutrition in Children.” *The American Journal of Clinical Nutrition*, vol. 75, no. 1, 1 Jan. 2002, pp. 92–98, academic.oup.com/ajcn/article/75/1/92/4689250, https://doi.org/10.1093/ajcn/75.1.92. Accessed 22 Nov. 2019.

5. Noaman, Eman, et al. “Vitamin E and Selenium Administration as a Modulator of Antioxidant Defense System.” *Biological Trace Element Research*, vol. 86, no. 1, 2002, pp. 55–64, https://doi.org/10.1385/bter:86:1:55. Accessed 14 Nov. 2022.

6. NHS. “Facts about Fat.” *Nhs.uk*, 23 Feb. 2022, www.nhs.uk/live-well/eat-well/food-types/different-fats-nutrition/#:~:text=Fat%20helps%20the%20body%20absorb.