Applying PCA to the USDA National Nutrient Database

Data Analysis Assignment Lesson 5

David Chen

Feburary 18th, 2020

Introduction

The USDA National Nutrient Database contains records of thousands of food items and their nutritional content. In order to better understand how the different nutrients are related to each other and how we can reduce the total number of variables, we will be conducting a Principal Component Analysis (PCA). After running this analyses, we observe that approximately 16-22 components are required to explain most of the variance, and that none of the first few components are particularly identifiable.

Data

For this analysis, we will be analyzing the USDA National Nutrient Database for Standard Reference (SR), release 28. This dataset was initially issued in August 2014, containing 8,789 food items and 150 different food components. For our analyses however, we will remove all NAs and duplicates, reducing the total number of food items to 2223. We will also only consider 46 of the nutrient variables, as we wish to conduct PCA on this data set, and that requires quantitative variables only.

For example, ‘Butter, with salt’ is an observation in this dataset, and nutrients such as water, protein, vitamin C, vitamin A, and cholesterol are included. While not all variables are in the same units, these differences will be accounted for later.

Exploratory Data Analysis

Examining Table 1, we can see that the means and variances differ substantially between all the variables. While the means range from `Copper` (0.172) to `Vit_A_IU` (499.490, the variances range from `Riboflavin` (0.087) to `Vit_A_IU` (5,098,719) . These vast differences will need to be accounted for prior to conducting principal component analysis for reasonable results.

Table 1: Table of Means and Variances

Food Item	Mean	Variance
Water_(g)	57.573	834.952
Energ_Kcal	210.855	25169.721
Protein_(g)	12.854	128.404
Lipid_Tot_(g)	10.153	206.014
Ash_(g)	1.868	15.963
Carbohydrt_(g)	17.553	607.600
Fiber_TD_(g)	1.763	16.222
Sugar_Tot_(g)	7.182	216.403
Calcium_(mg)	76.161	74738.686
Iron_(mg)	1.890	15.060
Magnesium_(mg)	33.580	3197.176
Phosphorus_(mg)	176.951	94438.343
Potassium_(mg)	298.560	260252.445
Sodium_(mg)	338.987	1918025.361
Zinc_(mg)	2.185	14.386
Copper_(mg)	0.172	0.158
Manganese_(mg)	0.459	10.761
Selenium_(µg)	17.079	1998.693
Vit_C_(mg)	6.356	2032.894
Thiamin_(mg)	0.175	0.127
Riboflavin_(mg)	0.212	0.087
Niacin_(mg)	3.317	12.925

Food Item	Mean	Variance
Panto_Acid_(mg)	0.586	0.963
Vit_B6_(mg)	0.257	0.098
Folate_Tot_(µg)	31.080	6855.954
Folic_Acid_(µg)	9.052	2064.629
Food_Folate_(µg)	22.011	4682.490
Folate_DFE_(µg)	37.396	10833.977
Choline_Tot_(mg)	46.626	5790.941
Vit_B12_(µg)	0.961	3.267
Vit_A_IU	499.490	5098719.551
Vit_A_RAE	48.207	35472.386
Retinol_(µg)	27.888	24111.535
Alpha_Carot_(µg)	26.795	52251.106
Beta_Carot_(µg)	225.179	1490031.925
Beta_Crypt_(µg)	10.922	24471.768
Lycopene_(µg)	162.429	2469441.007
Lut+Zea_(µg)	173.919	1226785.708
Vit_E_(mg)	0.999	10.507
Vit_D_µg	0.517	4.663
Vit_D_IU	20.684	7462.127
Vit_K_(µg)	13.271	5920.390
FA_Sat_(g)	3.396	34.016
FA_Mono_(g)	3.944	42.992
FA_Poly_(g)	1.910	19.958
Cholestrl_(mg)	42.198	11609.559

Conducting the PCA

We begin by conducting the principal component analysis. Since the original variables were on different scales (μg, mg, g), we center and scale all the variables.

From the scree plot in Figure 1, we can note that there is no distinct 'elbow' point where the proportion of variance explained by each component levels off. Since all the components explain a relatively low amount of variance, it is difficult to identify a clear cutoff point. This idea is further emphasized by Figure 2, where the variance explained by each component is a smooth, gradual increase. Thus,it is unclear how many principal components are necessary to adequately represent this data in a lower dimension. Depending on how we wish to use these components, we may wish to select between 16 (81% of variance explained) and 22 (91%) components.

Examining the factor loadings of the first 5 principal components (shown in Table 2 in the Appendix), none of them appear to provide an easily interpretable combination of nutrients. From the first four PCs, none of the individual loadings exceed an absolute value of 0.34, and so they are difficult to interpret. While technically speaking one could interpret PC5 as the absence of Ash (-0.47977465), Calcium (-0.42035926), and Phosphorus (-0.39606761), the sheer number of variables makes it difficult to derive any definitive conclusion.

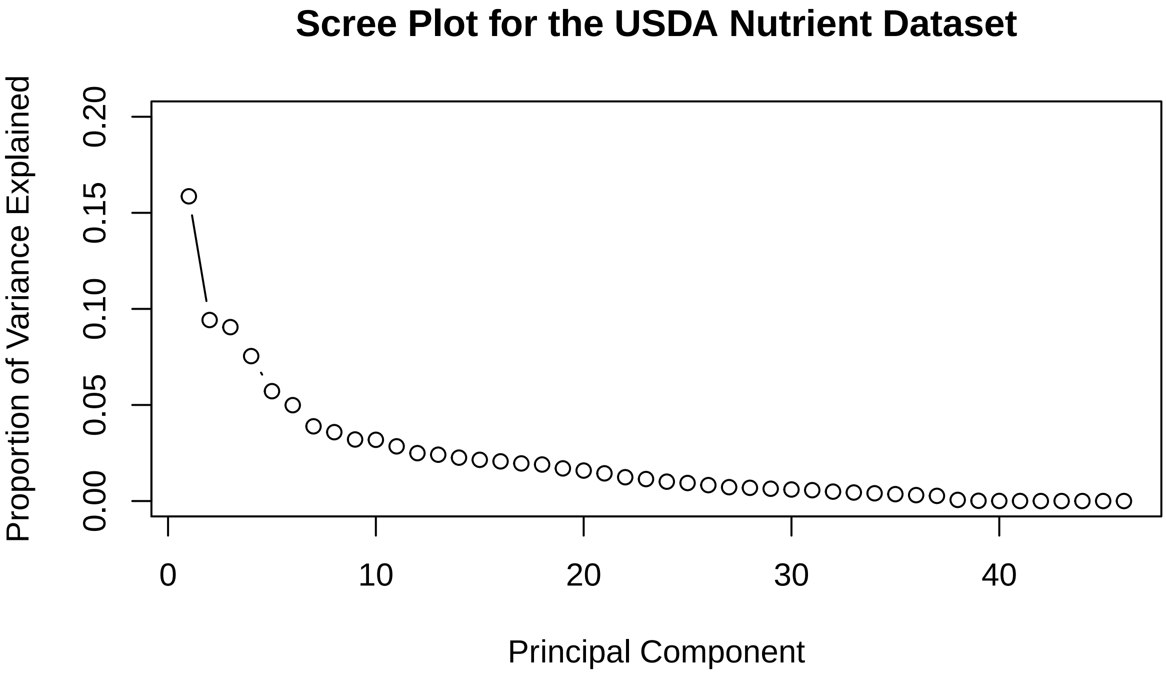

Figure 1: Scree Plot

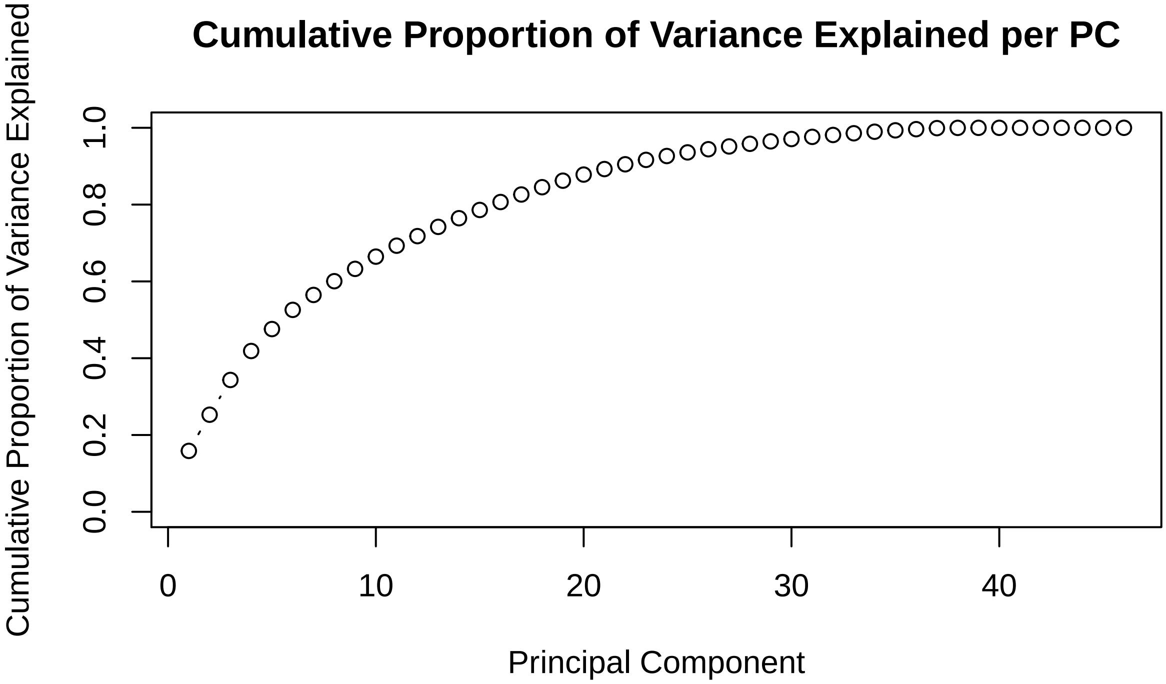

Figure 2: Cumulative Proportion of Variance Explained

Conclusion

After conducting a principal component analysis on the scaled USDA National Nutrient Database, we are able to reduce the total number of variables from 46 to 22 components (91% variance explained). While the individual components are not easily interpretable, this reduction of total variables can prove to be extremely beneficial for modeling, visualization, and data compression.

Appendix

Table 2: Factor Loadings

	PC1	PC2	PC3	PC4	PC5
Water_(g)	0.255	-0.031	0.276	-0.024	0.016
Energ_Kcal	-0.232	0.124	-0.304	-0.089	0.115
Protein_(g)	-0.138	0.263	0.188	-0.069	-0.060
Lipid_Tot_(g)	-0.161	0.185	-0.278	-0.241	0.149
Ash_(g)	-0.110	-0.026	-0.048	0.000	-0.480
Carbohydrt_(g)	-0.125	-0.185	-0.236	0.199	0.004
Fiber_TD_(g)	-0.188	-0.216	-0.107	0.029	-0.065
Sugar_Tot_(g)	-0.021	-0.116	-0.208	0.116	0.013
Calcium_(mg)	-0.113	-0.033	-0.033	-0.016	-0.420
Iron_(mg)	-0.221	-0.084	0.050	0.042	-0.086
Magnesium_(mg)	-0.246	-0.049	-0.092	-0.011	-0.128
Phosphorus_(mg)	-0.153	0.069	-0.002	-0.025	-0.396
Potassium_(mg)	-0.137	-0.051	0.014	-0.008	-0.262
Sodium_(mg)	-0.034	-0.006	-0.048	0.007	-0.307
Zinc_(mg)	-0.109	0.155	0.118	-0.053	-0.043
Copper_(mg)	-0.195	-0.016	-0.054	-0.006	-0.081
Manganese_(mg)	-0.137	-0.062	0.016	0.022	-0.032
Selenium_(µg)	-0.077	0.121	0.042	-0.052	-0.023
Vit_C_(mg)	-0.036	-0.103	0.047	0.028	0.006
Thiamin_(mg)	-0.215	-0.034	0.098	0.197	0.138
Riboflavin_(mg)	-0.237	0.020	0.160	0.044	0.010
Niacin_(mg)	-0.190	0.121	0.212	0.062	0.077
Panto_Acid_(mg)	-0.214	0.062	0.167	0.022	0.033
Vit_B6_(mg)	-0.181	0.058	0.235	0.000	0.044
Folate_Tot_(µg)	-0.247	-0.138	0.090	0.234	0.166
Folic_Acid_(µg)	-0.152	-0.081	0.061	0.191	0.159
Food_Folate_(µg)	-0.198	-0.113	0.068	0.156	0.095
Folate_DFE_(µg)	-0.243	-0.134	0.090	0.244	0.181
Choline_Tot_(mg)	-0.106	0.207	0.174	-0.136	-0.024
Vit_B12_(µg)	-0.072	0.197	0.234	-0.108	-0.018
Vit_A_IU	-0.036	-0.322	0.111	-0.337	0.047
Vit_A_RAE	-0.085	-0.191	0.132	-0.303	0.092
Retinol_(µg)	-0.092	0.003	0.096	-0.148	0.094
Alpha_Carot_(µg)	0.014	-0.178	0.054	-0.174	0.030
Beta_Carot_(µg)	-0.015	-0.329	0.091	-0.306	0.026
Beta_Crypt_(µg)	-0.047	-0.203	0.051	-0.224	0.000
Lycopene_(µg)	0.012	-0.039	-0.004	0.022	-0.037
Lut+Zea_(µg)	-0.016	-0.249	0.073	-0.195	0.000
Vit_E_(mg)	-0.182	-0.057	-0.111	-0.162	0.066
Vit_D_µg	-0.050	0.098	0.162	-0.063	-0.017
Vit_D_IU	-0.050	0.098	0.162	-0.063	-0.017
Vit_K_(µg)	-0.050	-0.198	0.027	-0.107	-0.057
FA_Sat_(g)	-0.086	0.168	-0.202	-0.196	0.110
FA_Mono_(g)	-0.143	0.177	-0.238	-0.215	0.140
FA_Poly_(g)	-0.157	0.069	-0.233	-0.149	0.097
Cholestrl_(mg)	-0.052	0.205	0.146	-0.150	0.012