Analyzing Food Consumption Using Principal Component Analysis

Principal Component Analysis (PCA) is a useful technique to summarize a large set of indicators into a smaller number of dimensions. To demonstrate how this method applies to a multivariate analysis, let's take a look at the Food Consumption dataset.

The table below represents the per capita food consumption of households across 31 provinces in China in 2018. The original data from the statistical yearbook divides the foods consumed by residents into 10 kinds of staples: Grain, Oil, Vegetables & Fungi, Meat, Poultry, Aquatic, Eggs, Diary, Fruits and Sugar.

Per Capita Annual Food Consumption in China (2018)

unit: kg

Region	Grain	Oil	Vege&Fungi	Meat	Poultry	Aquatic	Eggs	Dairy	Fruits	Sugar	Avg
Beijing	91.93	7.22	106.34	25.95	5.84	8.95	14.61	26.02	74.68	1.09	36.26
Tianjin	118.27	9.91	116.81	26.85	5.73	16.71	17.72	18.61	86.24	1.37	41.82
Hebei	130.84	7.64	95.53	23.03	4.89	5.88	13.69	14.45	66.82	1.08	36.39
Shanxi	137.64	7.46	83.69	15.31	2.42	2.62	11.28	15.66	55.93	1.04	33.30
Inner Mongolia	153.00	7.62	94.02	34.33	5.62	5.24	9.20	22.21	58.00	1.36	39.06
Liaoning	127.82	10.28	107.73	26.99	5.09	13.69	12.11	14.90	61.80	1.20	38.16
Jilin	132.64	10.75	92.26	24.78	4.88	7.96	10.10	10.02	53.92	1.27	34.86
Heilongjiang	139.88	12.88	95.79	25.70	5.35	9.03	10.86	10.39	64.32	1.75	37.59
Shanghai	110.53	8.04	103.60	31.44	12.35	24.49	11.99	20.78	63.52	1.53	38.83
Jiangsu	121.91	9.14	100.09	28.60	10.70	17.84	10.53	15.08	45.07	1.12	36.01
Zhejiang	132.85	11.56	91.72	29.76	10.72	22.88	8.43	13.19	52.59	1.54	37.52
Anhui	139.45	9.46	95.25	28.34	12.04	12.04	11.33	11.71	52.63	1.02	37.33
Fujian	125.10	9.14	90.83	34.72	11.53	23.95	8.65	11.75	47.19	1.69	36.45
Jiangxi	134.49	13.59	96.97	30.57	8.71	12.71	7.21	10.80	41.54	1.11	35.77
Shandong	117.21	7.59	92.79	23.89	6.04	12.13	16.00	16.45	74.22	0.91	36.72
Henan	123.41	8.38	84.73	18.48	6.00	3.98	12.80	12.54	56.03	1.26	32.76
Hubei	110.86	10.51	111.37	28.40	5.58	15.41	7.05	6.83	43.66	0.76	34.04
Hunan	137.64	11.75	94.36	34.54	10.81	11.89	7.77	6.65	57.48	1.27	37.42
Guangdong	108.67	9.16	100.57	40.99	21.08	22.03	7.39	8.63	39.45	1.56	35.95
Guangxi	128.28	8.06	85.16	35.73	18.53	9.85	5.47	5.57	38.96	1.21	33.68
Hainan	94.84	8.84	90.68	33.93	18.76	27.02	4.79	4.67	28.36	1.06	31.29
Chongqing	135.74	13.85	132.01	43.86	10.01	9.93	9.77	12.64	43.68	2.77	41.42
Sichuan	146.58	12.07	120.75	45.44	10.45	7.20	8.43	12.48	40.56	1.93	40.59
Guizhou	111.27	6.64	75.21	31.73	4.65	2.21	3.38	4.38	33.19	0.79	27.34
Yunnan	117.31	7.12	83.85	31.93	6.69	3.52	4.29	5.13	29.62	1.12	29.06
Tibet	208.90	15.93	42.71	29.45	1.50	0.47	3.64	14.32	5.99	4.35	32.73
Shaanxi	131.68	10.65	83.53	16.24	2.90	2.79	7.80	13.76	46.64	1.03	31.70
Gansu	151.84	9.42	79.93	20.19	4.81	2.43	8.68	13.65	75.62	1.89	36.85
Qinghai	113.19	9.26	52.35	27.02	2.92	1.69	3.72	17.63	24.83	1.45	25.41
Ningxia	112.12	7.23	87.53	16.26	6.75	2.54	5.92	13.53	78.72	1.42	33.20
Xinjiang	156.19	14.07	91.83	24.37	4.88	2.92	5.58	19.85	52.26	1.17	37.31

Source: National Bureau of Statistics

Given the information above, the goal of the analysis is to identify food consumption levels across all the regions. However, to elaborate as much variance as possible, we

need all the ten variables. Therefore, our usual practice is to compute the average: adding up the numbers together within each province and divide it by the number of food types. At first glance averaging seems like a reasonable solution, but it is hard to tell the difference when average scores for some regions are close with data heavily skewed towards different food categories. For instance, people in Hainan and Shaanxi Province both consume around 31kg last year, while grain remains the main food in Shaanxi and residents from Hainan have a relatively balanced diet. In this case, PCA performs better than using a single index of average or sum.

Selecting Number of Principal Components

What PCA does is to find out one or more of most effective ways to differentiate the regions to the maximum extent. During the process of the analysis we use more than one principal components (PCs) weighted by original 10 variables to measure the food consumption levels of regions, and these PCs will be ordered by their importance, the ability or power to differentiate regions. By doing this, PCA will help us find a smaller number of components that captures as much information as possible from the original dataset.

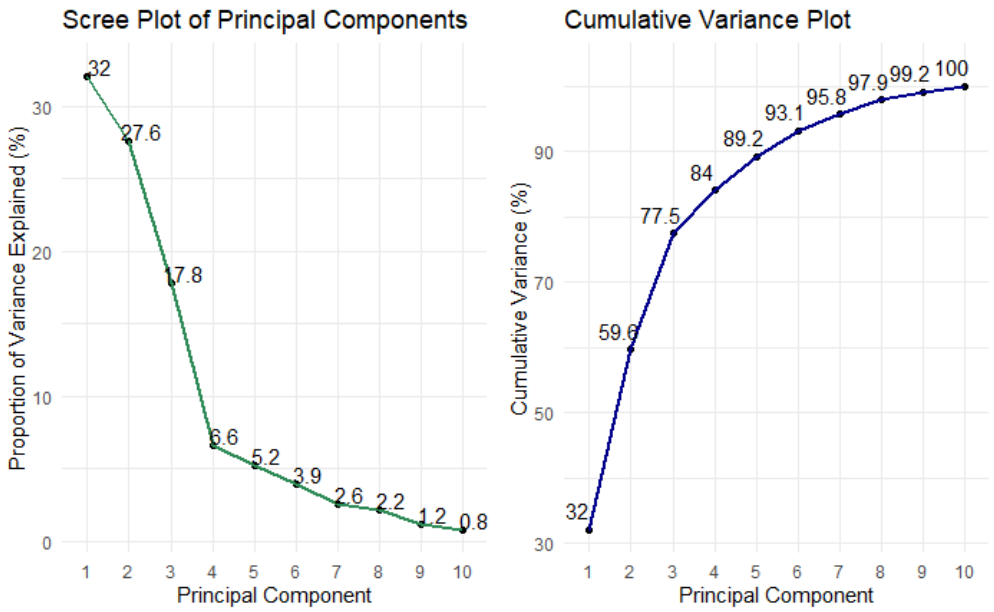

Figure 1

To see how much of information from the original dataset is explained by each of the principal component, we need to know the proportion of variance explained (PVE) by each principal component. The figures above show that the first principal component (PC1) contributes to the most variance (32%) in the original dataset. As we move from the first principal component to the last one, the amount of variance explained decreases

while cumulative explained variance approaches 100%. To reduce the number of dimensions, we can choose the number of PCs based on the cumulative explained variance from the right plot. The first four principal components collectively explain 84% of variance in the data, which means we can reduce dimensionality from 10 to 4 while losing about 16% of variance.

Understanding Relationship between Variables and Principal Components

The heatmap below illustrates how each kind of food contribute to the principal components using correlation between each variable and principal component[1]. In general, grain, oil and vegetables are three critical kinds of foods to focus on, and the significance of other variables fades gradually. It can be easily seen that Grain and Oil are positively correlated with PC1, so PC1 roughly represents the consumption of traditional staple food. The main focus of PC2 is on meat, poultry and aquatic as they contribute the most amount of variance in PC2, while grain is negatively associated with this component. PC3 and PC4 move the attention to oil and diary respectively from other staples.

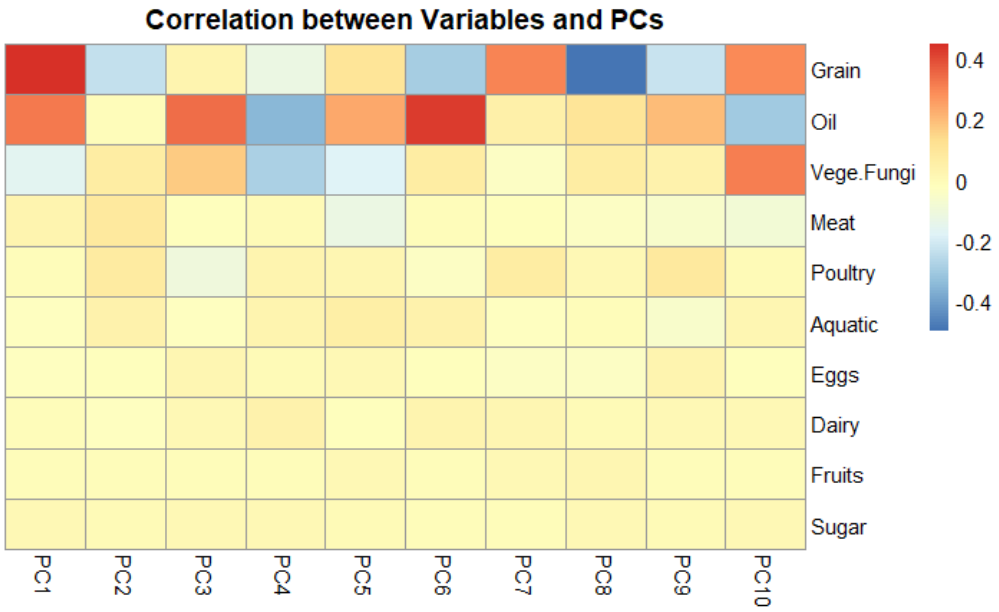

Figure 2

[1] Gaurav Kaushik, Visualizing Tools For Feature Importance and Principal Component Analysis, 2018, https://medium.com/cascade-bio-blog/creating-visualizations-to-better-understand-your-data-and-models-part-1-a51e7e5af9c0

Positions of Regions on Principal Components

The two plots below (Figure3 and Figure4) include the position of regions in terms of the first four principal components and scores on each axis appear in varying shades of blue. It can be seen that regions with similar dietary structure are close to each other. For instance, Sichuan and Chongqing both score high on the first three principals (corresponding to grain and meat) and low on the fourth principal (oil and dairy). In addition, Tibet scores highest on PC1 but low on PC2, reflecting relative food (particularly meat protein) scarcity in this area. Consumers from coastal regions such as Guangdong and Hainan spend far more on poultry and aquatic products and correspondingly score high on PC2. Meanwhile, people from developed regions (Beijing, Shanghai) pay more attention to nourishment and food variety, as reflected in the high scores on PC3 and PC4 and low scores on PC1 (corresponding to traditional staple food). Therefore, the government should take measures to improve the diet structure, especially the diet quality in less developed regions.

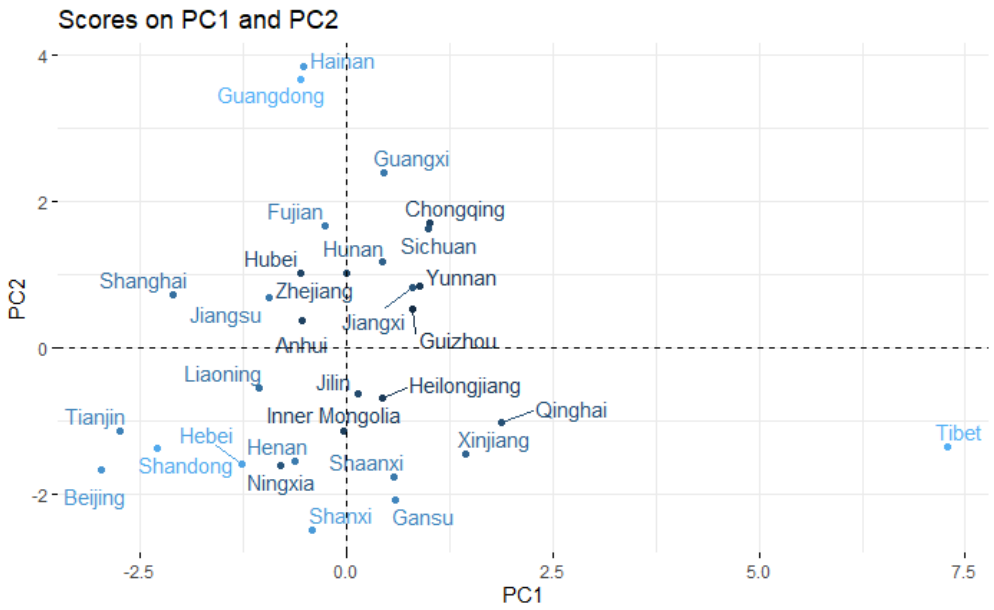

Figure 3

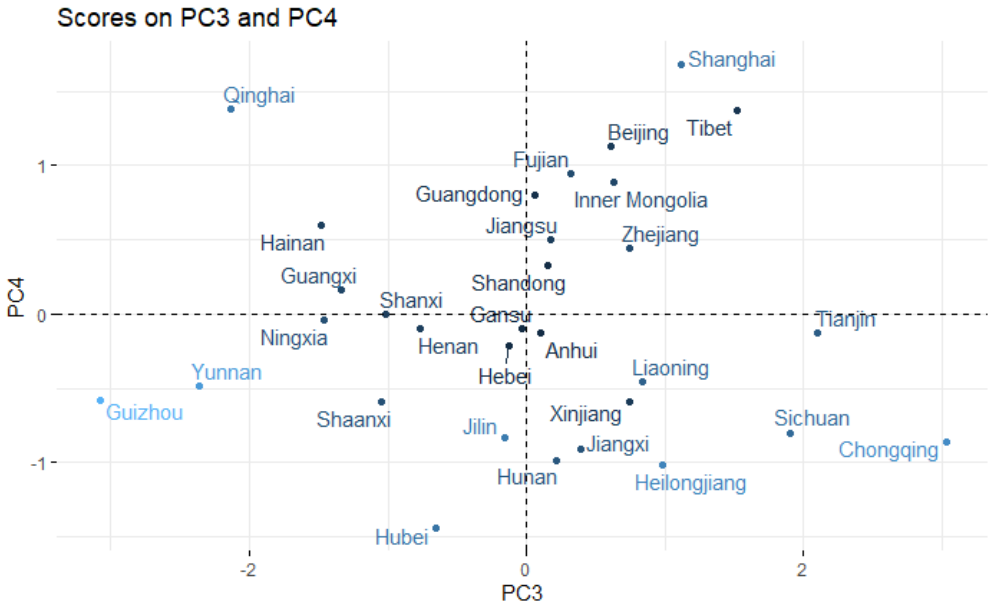

Figure 4

Principal Component Analysis reduces the computation costs and improves the interpretability in the process of evaluating the per capita food consumption. With fewer variables, it has become possible to train a predictive model as our next step of analysis.