| Principal Components | | |
|---|---|---|
| Feature\Component | 1 | 2 |
| calories | –0.847053 | –0.531508 |
| rating | 0.5315077 | –0.847053 |

| Variances | | |
|---|---|---|
| | 1 | 2 |
| Variance | 498.02448 | 78.932739 |
| Variance % | 86.319135 | 13.680865 |
| Cumulative Variance % | 86.319135 | 100 |

**FIGURE 4.8** OUTPUT FROM PRINCIPAL COMPONENTS ANALYSIS OF *CALORIES* AND *RATING*

is the projection onto $z_2$ using the weights $(-0.532, -0.847)$. For example, the first score for the 100% Bran cereal (with 70 calories and a rating of 68.4) is $(-0.847)(70-106.88)+(0.532)(68.4-42.67) = 44.92$. Note that the means of the new variables $z_1$ and $z_2$ are zero, since we've subtracted the mean of each variable. The sum of the variances $\text{var}(z_1) + \text{var}(z_2)$ is equal to the sum of the variances of the original variables, $\text{var}(calories) + \text{var}(rating)$. Furthermore the variances of $z_1$ and $z_2$ are 498 and 79, respectively, so the first principal component, $z_1$, accounts for 86% of the total variance. Since it captures most of the variability in the data, it seems reasonable to use one variable, the first principal score, to represent the two variables in the original data. Next, we generalize these ideas to more than two variables.

## Principal Components

Let us formalize the procedure described above so that it can easily be generalized to $p > 2$ variables. Denote by $X_1, X_2, \ldots, X_p$ the original $p$ variables. In PCA we are looking for a set of new variables $Z_1, Z_2, \ldots, Z_p$ that are weighted averages of the original variables (after subtracting their mean):

$$Z_i = a_{i,1}(X_1 - \bar{X}_1) + a_{i,2}(X_2 - \bar{X}_2) + \cdots + a_{i,p}(X_p - \bar{X}_p), \quad i = 1, \ldots, p, \tag{4.1}$$

where each pair of $Z$'s has correlation $= 0$. We then order the resulting $Z$'s by their variance, with $Z_1$ having the largest variance and $Z_p$ having the smallest variance. The software computes the weights $a_{i,j}$, which are then used in computing the principal component scores.

A further advantage of the principal components compared to the original data is that they are uncorrelated (correlation coefficient = 0). If we construct regression models using these principal components as independent variables, we will not encounter problems of multicollinearity.

| Scores | | |
|---|---|---|
| Pattern\Component | 1 | 2 |
| 100% Bran | 44.921528 | –2.197183 |
| 100% Natural Bran | –15.72526 | 0.3824165 |
| All-Bran | 40.149935 | 5.4072123 |
| All-Bran with Extra Fiber | 75.310772 | –12.99913 |
| Almond Delight | –7.041508 | 5.3576857 |
| Apple Cinnamon Cheerios | –9.632769 | 9.4873273 |
| Apple Jacks | –7.685031 | 6.3832549 |
| Basic 4 | –22.57211 | –7.520309 |
| Bran Chex | 17.731545 | 3.5061586 |
| Bran Flakes | 19.960454 | –0.046011 |
| Cap'n'Crunch | –24.19794 | 13.88515 |
| Cheerios | 1.6646701 | –8.517182 |
| Cinnamon Toast Crun | –23.25147 | 12.376784 |
| Clusters | –3.844296 | 0.2623499 |
| Cocoa Puffs | –13.23272 | 15.224501 |
| Corn Chex | –3.288971 | –0.622661 |
| Corn Flakes | 7.5299271 | 0.949875 |

**FIGURE 4.9** **PRINCIPAL SCORES FROM PRINCIPAL COMPONENTS ANALYSIS OF *CALORIES* AND *RATING* FOR THE FIRST 17 CEREALS.**

Let us return to the breakfast cereal dataset with all 15 variables, and apply PCA to the 13 numerical variables. The resulting output is shown in Figure 4.10. For simplicity, we removed three cereals that contained missing values.

| Principal Components | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Feature\Component | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| calories | –0.078 | 0.009 | 0.629 | 0.601 | 0.455 | 0.119 | –0.094 | 0.026 | –0.009 | 0.065 | –0.009 | –0.004 | –0.042 |
| protein | 0.001 | –0.009 | 0.001 | –0.003 | 0.056 | 0.113 | –0.258 | –0.655 | 0.202 | –0.256 | 0.045 | 0.005 | 0.616 |
| fat | 0.000 | –0.003 | 0.016 | 0.025 | –0.016 | –0.132 | –0.373 | 0.118 | –0.124 | –0.841 | 0.062 | 0.009 | –0.318 |
| sodium | –0.980 | –0.141 | –0.136 | 0.001 | 0.014 | 0.023 | –0.005 | –0.001 | 0.004 | –0.001 | 0.000 | 0.000 | –0.010 |
| fiber | 0.005 | –0.031 | –0.018 | –0.020 | 0.014 | 0.263 | –0.043 | 0.659 | –0.227 | –0.144 | –0.021 | –0.001 | 0.648 |
| carbo | –0.017 | 0.017 | 0.017 | –0.026 | 0.349 | –0.538 | 0.672 | –0.006 | 0.025 | –0.300 | 0.042 | –0.014 | 0.206 |
| sugars | –0.003 | 0.000 | 0.098 | 0.115 | –0.299 | 0.648 | 0.567 | –0.103 | 0.117 | –0.320 | 0.028 | –0.019 | –0.136 |
| potass | 0.135 | –0.987 | 0.037 | 0.042 | –0.047 | –0.050 | 0.018 | –0.015 | –0.001 | 0.006 | 0.000 | –0.001 | –0.006 |
| vitamins | –0.094 | –0.017 | 0.692 | –0.714 | –0.037 | 0.016 | –0.012 | –0.004 | –0.012 | 0.001 | 0.002 | –0.001 | –0.010 |
| shelf | 0.002 | –0.004 | 0.012 | –0.006 | –0.008 | –0.060 | –0.092 | 0.328 | 0.935 | –0.046 | –0.068 | 0.010 | 0.000 |
| weight | –0.001 | –0.001 | 0.004 | 0.003 | 0.003 | 0.009 | 0.024 | 0.003 | –0.002 | 0.006 | 0.093 | 0.995 | 0.000 |
| cups | –0.001 | 0.002 | 0.001 | –0.001 | 0.002 | –0.010 | 0.020 | –0.062 | –0.054 | –0.080 | –0.989 | 0.092 | 0.000 |
| rating | 0.075 | –0.072 | –0.308 | –0.335 | 0.758 | 0.413 | –0.018 | –0.012 | 0.036 | –0.023 | –0.003 | –0.003 | –0.188 |

| Variances | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| Variance | 7016.42 | 5028.83 | 512.739 | 367.93 | 70.951 | 4.3751 | 2.888 | 0.6078 | 0.4327 | 0.13722 | 0.0348 | 0.004 | 0.0000 |
| Variance % | 53.950 | 38.667 | 3.943 | 2.829 | 0.546 | 0.034 | 0.022 | 0.005 | 0.003 | 0.001 | 0.000 | 0.000 | 0.000 |
| Cumulative Variance % | 53.95 | 92.62 | 96.56 | 99.39 | 99.93 | 99.97 | 99.99 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |

**FIGURE 4.10** **PCA OUTPUT USING ALL 13 NUMERICAL VARIABLES IN THE BREAKFAST CEREALS DATASET.**

Note that the first three components account for more than 96% of the total variation associated with all 13 of the original variables. This suggests that we can capture most of the variability in the data with less than 25% of the number of original dimensions in the data. In fact, the first two principal components alone capture 92.6% of the total variation. However, these results are influenced by the scales of the variables, as we describe next.

## Normalizing the Data

A further use of PCA is to understand the structure of the data. This is done by examining the weights to see how the original variables contribute to the different principal components. In our example, it is clear that the first principal component is dominated by the sodium content of the cereal: it has the highest (in this case, positive) weight. This means that the first principal component is in fact measuring how much sodium is in the cereal. Similarly, the second principal component seems to be measuring the amount of potassium. Since both variables are measured in milligrams, whereas the other nutrients are measured in grams, the scale is obviously leading to this result. The variances of potassium and sodium are much larger than the variances of the other variables, and thus the total variance is dominated by these two variances. A solution is to normalize the data before performing the PCA. Normalization (or standardization) means replacing each original variable by a standardized version of the variable that has unit variance. This is easily accomplished by dividing each variable by its standard deviation. The effect of this normalization (standardization) is to give all variables equal importance in terms of the variability.

When should we normalize the data like this? It depends on the nature of the data. When the units of measurement are common for the variables (e.g., dollars), and when their scale reflects their importance (sales of jet fuel, sales of heating oil, etc.), it is probably best not to normalize (i.e., not to rescale the data so that they have unit variance). If the variables are measured in different units so that it is unclear how to compare the variability of different variables (e.g., dollars for some, parts per million for others) or if for variables measured in the same units, scale does not reflect importance (earnings per share, gross revenues), it is generally advisable to normalize. In this way the changes in units of measurement do not change the principal components' weights. In the rare situations where we can give relative weights to variables, we multiply the normalized variables by these weights before doing the principal components analysis.

Thus far we have calculated principal components using the covariance matrix. An alternative to normalizing and then performing PCA is to perform PCA on the correlation matrix instead of the covariance matrix. Most software programs allow the user to choose between the two. Remember, using the correlation matrix means that you are operating on the normalized data.

Returning to the breakfast cereal data, we normalize the 13 variables due to the different scales of the variables and then perform PCA (or equivalently, we use PCA applied to the correlation matrix). The output is shown in Figure 4.11. Now we find that we need 7 principal components to account for

| Principal Components | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Feature\Component | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| calories | –0.300 | 0.393 | 0.115 | –0.204 | 0.204 | –0.256 | 0.026 | 0.002 | –0.030 | 0.500 | –0.214 | 0.492 | –0.234 |
| protein | 0.307 | 0.165 | 0.277 | –0.301 | 0.320 | 0.121 | –0.283 | 0.427 | –0.535 | –0.022 | 0.032 | –0.100 | 0.186 |
| fat | –0.040 | 0.346 | –0.205 | –0.187 | 0.587 | 0.348 | 0.051 | –0.063 | 0.460 | –0.145 | –0.067 | –0.291 | –0.090 |
| sodium | –0.183 | 0.137 | 0.389 | –0.120 | –0.338 | 0.664 | 0.284 | –0.177 | –0.215 | –0.001 | –0.087 | –0.054 | –0.239 |
| fiber | 0.453 | 0.180 | 0.070 | –0.039 | –0.255 | 0.064 | –0.112 | –0.216 | 0.244 | 0.295 | –0.531 | –0.059 | 0.442 |
| carbo | –0.192 | –0.149 | 0.562 | –0.088 | 0.183 | –0.326 | 0.260 | –0.167 | 0.117 | 0.241 | 0.179 | –0.476 | 0.225 |
| sugars | –0.228 | 0.351 | –0.355 | 0.023 | –0.315 | –0.152 | –0.228 | 0.063 | –0.225 | 0.252 | –0.003 | –0.614 | –0.167 |
| potass | 0.402 | 0.301 | 0.068 | –0.091 | –0.148 | 0.025 | –0.149 | –0.262 | 0.167 | 0.177 | 0.729 | 0.121 | –0.128 |
| vitamins | –0.116 | 0.173 | 0.388 | 0.604 | –0.049 | 0.129 | –0.294 | 0.457 | 0.346 | 0.052 | 0.019 | 0.000 | –0.060 |
| shelf | 0.171 | 0.265 | –0.002 | 0.639 | 0.329 | –0.052 | 0.175 | –0.414 | –0.416 | –0.046 | –0.059 | –0.017 | 0.000 |
| weight | –0.050 | 0.450 | 0.247 | –0.153 | –0.221 | –0.399 | –0.014 | –0.075 | 0.065 | –0.692 | –0.113 | 0.031 | 0.000 |
| cups | –0.295 | –0.212 | 0.140 | –0.047 | 0.121 | 0.099 | –0.749 | –0.499 | –0.050 | –0.077 | –0.055 | 0.032 | 0.000 |
| rating | 0.438 | –0.252 | 0.182 | –0.038 | 0.058 | –0.186 | –0.063 | –0.015 | 0.063 | 0.012 | –0.276 | –0.189 | –0.743 |

| Variances | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| Variance | 3.634 | 3.148 | 1.909 | 1.019 | 0.989 | 0.722 | 0.672 | 0.416 | 0.316 | 0.092 | 0.063 | 0.019 | 0.000 |
| Variance % | 27.95 | 24.22 | 14.69 | 7.84 | 7.61 | 5.55 | 5.17 | 3.20 | 2.43 | 0.71 | 0.49 | 0.15 | 0.00 |
| Cumulative Variance % | 27.95 | 52.17 | 66.85 | 74.70 | 82.31 | 87.86 | 93.03 | 96.23 | 98.66 | 99.36 | 99.85 | 100.00 | 100.00 |

**FIGURE 4.11** **PCA OUTPUT USING ALL *NORMALIZED* 13 NUMERICAL VARIABLES IN THE BREAKFAST CEREALS DATASET.**

more than 90% of the total variability. The first 2 principal components account for only 52% of the total variability, and thus reducing the number of variables to 2 would mean losing a lot of information. Examining the weights, we see that the first principal component measures the balance between 2 quantities: (1) calories and cups (large negative weights) vs. (2) protein, fiber, potassium, and consumer rating (large positive weights). High scores on principal component 1 mean that the cereal is high in calories and the amount per bowl, and low in protein and potassium. Unsurprisingly, this type of cereal is associated with a low consumer rating. The second principal component is most affected by the weight of a serving, and the third principal component by the carbohydrate content. We can continue labeling the next principal components in a similar fashion to learn about the structure of the data.

When the data can be reduced to two dimensions, a useful plot is a scatter plot of the first vs. second principal scores with labels for the observations (if the dataset is not too large). To illustrate this, Figure 4.12 displays the first two principal component scores for the breakfast cereals.

We can see that as we move from right (bran cereals) to left, the cereals are less "healthy" in the sense of high calories, low protein and fiber, and so on. Also, moving from bottom to top, we get heavier cereals (moving from puffed rice to raisin bran). These plots are especially useful if interesting clusterings of