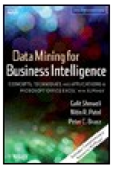


Data Mining for Business Intelligence

Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner, Second Edition

by Galit Shmueli, Nitin R. Patel and Peter C. Bruce

Chapter 4: Dimension Reduction

In this chapter we describe the important step of dimension reduction. The dimension of a dataset, which is the number of variables, must be reduced for the data mining algorithms to operate efficiently. We present and discuss several dimension reduction approaches: (1) Incorporating domain knowledge to remove or combine categories, (2) using data summaries to detect information overlap between variables (and remove or combine redundant variables or categories), (3) using data conversion techniques such as converting categorical variables into numerical variables, and (4) employing automated reduction techniques, such as principal components analysis (PCA), where a new set of variables (which are weighted averages of the original variables) is created. These new variables are uncorrelated and a small subset of them usually contains most of their combined information (hence, we can reduce dimension by using only a subset of the new variables). Finally, we mention data mining methods such as regression models and regression and classification trees, which can be used for removing redundant variables and for combining "similar" categories of categorical variables.

4.1 Introduction

In data mining one often encounters situations where there are a large number of variables in the database. In such situations it is very likely that subsets of variables are highly correlated with each other. Included in a classification or prediction model, highly correlated variables, or variables that are unrelated to the outcome of interest, can lead to overfitting, and accuracy and reliability can suffer. Large numbers of variables also pose computational problems for some models (aside from questions of correlation). In model deployment, superfluous variables can increase costs due to the collection and processing of these variables. The *dimensionality* of a model is the number of independent or input variables used by the model. One of the key steps in data mining, therefore, is finding ways to reduce dimensionality without sacrificing accuracy. In the artificial intelligence literature, dimension reduction is often referred to as *factor selection* or *feature extraction*.

4.2 Practical Considerations

Although data mining prefers automated methods over domain knowledge, it is important at the first step of data exploration to make sure that the variables measured are reasonable for the task at hand. The integration of expert knowledge through a discussion with the data provider (or user) will probably lead to better results. Practical considerations include: Which variables are most important for the task at hand, and which are most likely to be useless? Which variables are likely to contain much error? Which variables will be available for measurement (and what will it cost to measure them) in the future if the analysis is repeated? Which variables can actually be measured before the outcome occurs? (For example, if we want to predict the closing price of an ongoing online auction, we cannot use the number of bids as a predictor because this will not be known until the auction closes.)

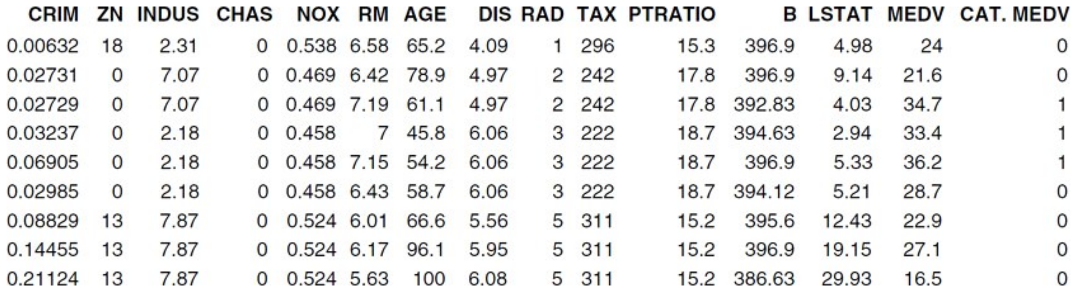
Example 1: House Prices in Boston

We return to the Boston housing example introduced in Chapter 2. For each neighborhood, a number of variables are given, such as the crime rate, the student/teacher ratio, and the median value of a housing unit in the neighborhood. A description of all 14 variables is given in Table 4.1. The first 10 records of the data are shown in Figure 4.1. The first row in this figure represents the first neighborhood, which had an average per capita crime rate of 0.006, 18% of the residential land zoned for lots over 25, 000 ft^2, 2.31% of the land devoted to nonretail business, no border on the Charles River, and so on.

Table 4.1: DESCRIPTION OF VARIABLES IN THE BOSTON HOUSING DATASET

CRIM	Crime rate
ZN	Percentage of residential land zoned for lots over 25, 000 ft^2
INDUS	Percentage of land occupied by nonretail business
CHAS	Charles River dummy variable (= 1 if tract bounds river; =0 otherwise)
NOX	Nitric oxide concentration (parts per 10 million)
RM	Average number of rooms per dwelling
AGE	Percentage of owner-occupied units built prior to 1940
DIS	Weighted distances to five Boston employment centers

RAD	Index of accessibility to radial highways
TAX	Full-value property tax rate per $10, 000
PTRATIO	Pupil/teacher ratio by town
B	1000(Bk minus 0.63)^2, where Bk is the proportion of blacks by town
LSTAT	% Lower status of the population
MEDV	Median value of owner-occupied homes in $1000s



CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV	CAT. MEDV
0.00632	18	2.31	0	0.538	6.58	65.2	4.09	1	296	15.3	396.9	4.98	24	0
0.02731	0	7.07	0	0.469	6.42	78.9	4.97	2	242	17.8	396.9	9.14	21.6	0
0.02729	0	7.07	0	0.469	7.19	61.1	4.97	2	242	17.8	392.83	4.03	34.7	1
0.03237	0	2.18	0	0.458	7	45.8	6.06	3	222	18.7	394.63	2.94	33.4	1
0.06905	0	2.18	0	0.458	7.15	54.2	6.06	3	222	18.7	396.9	5.33	36.2	1
0.02985	0	2.18	0	0.458	6.43	58.7	6.06	3	222	18.7	394.12	5.21	28.7	0
0.08829	13	7.87	0	0.524	6.01	66.6	5.56	5	311	15.2	395.6	12.43	22.9	0
0.14455	13	7.87	0	0.524	6.17	96.1	5.95	5	311	15.2	396.9	19.15	27.1	0
0.21124	13	7.87	0	0.524	5.63	100	6.08	5	311	15.2	386.63	29.93	16.5	0

FIGURE 4.1: FIRST NINE RECORDS IN THE BOSTON HOUSING DATASET

4.3 Data Summaries

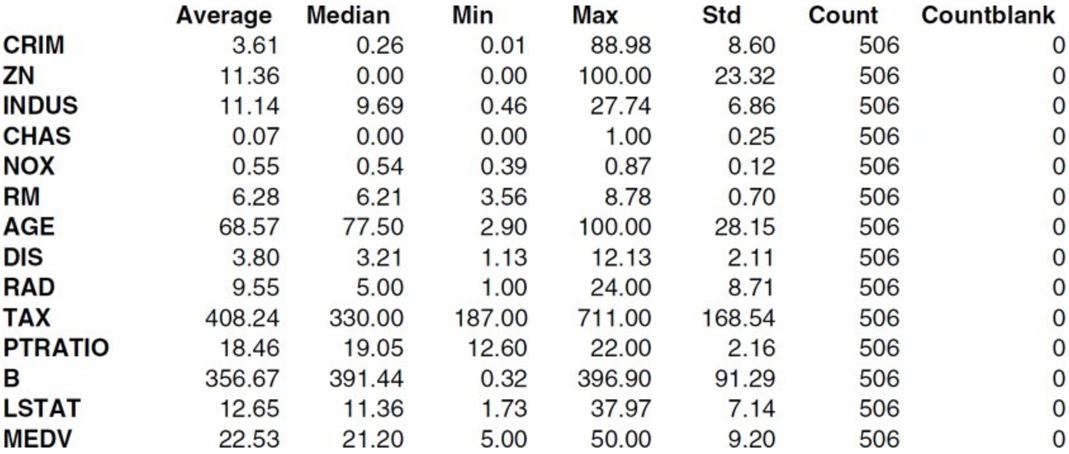
As we have seen in Chapter 3 on data visualization, an important initial step of data exploration is getting familiar with the data and their characteristics through summaries and graphs. The importance of this step cannot be overstated. The better you understand the data, the better the results from the modeling or mining process will be.

Numerical summaries and graphs of the data are very helpful for data reduction. The information that they convey can assist in combining categories of a categorical variable, in choosing variables to remove, in assessing the level of information overlap between variables, and more. Before discussing such strategies for reducing the dimension of a data set, let us consider useful summaries and tools.

Summary Statistics

Excel has several functions and facilities that assist in summarizing data. The functions *average, stdev, min, max, median*, and *count* are very helpful for learning about the characteristics of each variable. First, they give us information about the scale and type of values that the variable takes. The min and max functions can be used to detect extreme values that might be errors. The average and median give a sense of the central values of that variable, and a large deviation between the two also indicates skew. The standard deviation gives a sense of how dispersed the data are (relative to the mean). Other functions, such as *countblank*, which gives the number of empty cells, can tell us about missing values. It is also possible to use Excel's *Descriptive Statistics* facility in the *Data > Data Analysis* menu (in Excel 2003: *Tools > Data Analysis)*. This will generate a set of 13 summary statistics for each of the variables.

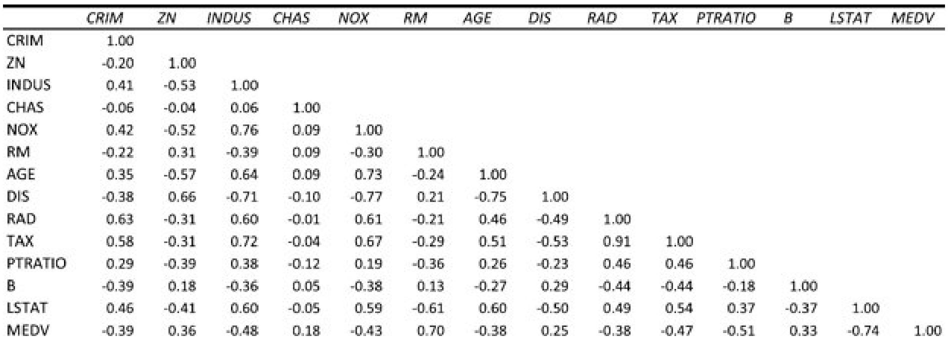
Figure 4.2 shows six summary statistics for the Boston housing example. We see immediately that the different variables have very different ranges of values. We will see soon how variation in scale across variables can distort analyses ifnot treated properly. Another observation that can be made is that the average of the first variable, CRIM (as well as several others), is much larger than the median, indicating right skew. None of the variables have empty cells. There also do not appear to be indications of extreme values that might result from typing errors.



	Average	Median	Min	Max	Std	Count	Countblank
CRIM	3.61	0.26	0.01	88.98	8.60	506	0
ZN	11.36	0.00	0.00	100.00	23.32	506	0
INDUS	11.14	9.69	0.46	27.74	6.86	506	0
CHAS	0.07	0.00	0.00	1.00	0.25	506	0
NOX	0.55	0.54	0.39	0.87	0.12	506	0
RM	6.28	6.21	3.56	8.78	0.70	506	0
AGE	68.57	77.50	2.90	100.00	28.15	506	0
DIS	3.80	3.21	1.13	12.13	2.11	506	0
RAD	9.55	5.00	1.00	24.00	8.71	506	0
TAX	408.24	330.00	187.00	711.00	168.54	506	0
PTRATIO	18.46	19.05	12.60	22.00	2.16	506	0
B	356.67	391.44	0.32	396.90	91.29	506	0
LSTAT	12.65	11.36	1.73	37.97	7.14	506	0
MEDV	22.53	21.20	5.00	50.00	9.20	506	0

FIGURE 4.2: SUMMARY STATISTICS FOR THE BOSTON HOUSING DATA

Next, we summarize relationships between two or more variables. For numerical variables, we can compute pairwise correlations (using the Excel function *correl)*. We can also obtain a complete matrix of correlations between each pair of variables in the data using Excel's *Correlation* facility in the *Data* > *Data Analysis* menu (in Excel 2003, *Tools* > *Data Analysis)*. Figure 4.3 shows the correlation matrix for a subset of the Boston housing variables. We see that most are low and that many are negative. Recall also the visual display of a correlation matrix via a heatmap (see Figure 3.3 for the heatmap corresponding to this correlation table). We will return to the importance of the correlation matrix soon, in the context of correlation analysis.



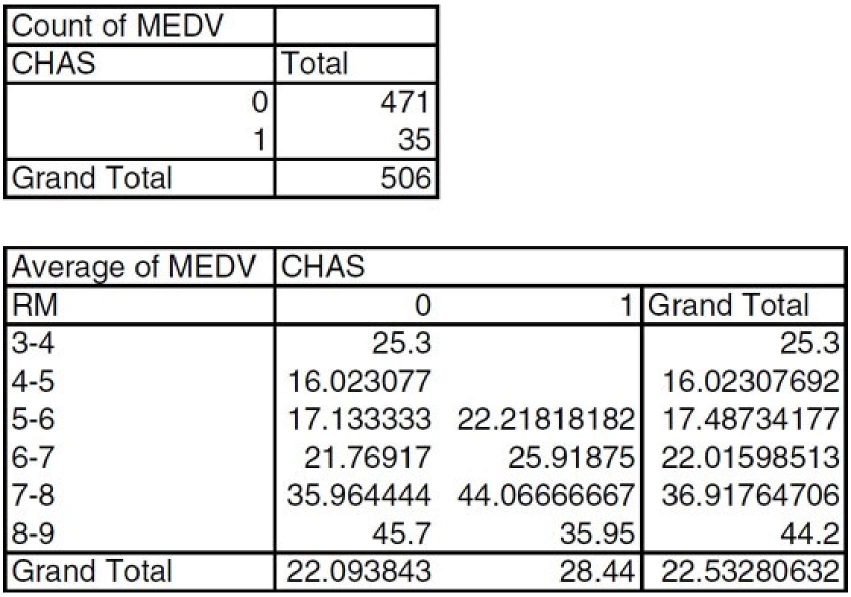
	CRIM	*ZN*	*INDUS*	*CHAS*	*NOX*	*RM*	*AGE*	*DIS*	*RAD*	*TAX*	*PTRATIO*	*B*	*LSTAT*	*MEDV*
CRIM	1.00													
ZN	-0.20	1.00												
INDUS	0.41	-0.53	1.00											
CHAS	-0.06	-0.04	0.06	1.00										
NOX	0.42	-0.52	0.76	0.09	1.00									
RM	-0.22	0.31	-0.39	0.09	-0.30	1.00								
AGE	0.35	-0.57	0.64	0.09	0.73	-0.24	1.00							
DIS	-0.38	0.66	-0.71	-0.10	-0.77	0.21	-0.75	1.00						
RAD	0.63	-0.31	0.60	-0.01	0.61	-0.21	0.46	-0.49	1.00					
TAX	0.58	-0.31	0.72	-0.04	0.67	-0.29	0.51	-0.53	0.91	1.00				
PTRATIO	0.29	-0.39	0.38	-0.12	0.19	-0.36	0.26	-0.23	0.46	0.46	1.00			
B	-0.39	0.18	-0.36	0.05	-0.38	0.13	-0.27	0.29	-0.44	-0.44	-0.18	1.00		
LSTAT	0.46	-0.41	0.60	-0.05	0.59	-0.61	0.60	-0.50	0.49	0.54	0.37	-0.37	1.00	
MEDV	-0.39	0.36	-0.48	0.18	-0.43	0.70	-0.38	0.25	-0.38	-0.47	-0.51	0.33	-0.74	1.00

FIGURE 4.3: CORRELATION TABLE FOR BOSTON HOUSING DATA, GENERATED USING EXCEL'S DATA ANALYSIS MENU

Pivot Tables

Another very useful tool is Excel's *pivot tables, inthe Insert* > *Data* menu (in Excel 2003, in the *Data* menu). These are interactive tables that can combine information from multiple variables and compute a range of summary statistics (count, average, percentage, etc.). A simple example is the average MEDV for neighborhoods that bound the Charles River versus those that do not. First, we get a count of neighborhoods bordering the river. The Excel pivot table in Figure 4.4 (top panel) was obtained by selecting CHAS as a "row labels" field and MEDV or any other variable as a "values" field, using the "count" summary. It appears that the majority of neighborhoods (471 of 506) do not bound the river. By double-clicking on a certain cell, the complete data for records in that cell are shown on a new worksheet. For instance, double-clicking on the cell containing 471 will display the complete records of neighborhoods that do not bound the river.

Pivot tables can be used for multiple variables. For categorical variables we obtain a breakdown of the records by the combination of categories. For instance, the bottom panel of Figure 4.4 shows the average MEDV by CHAS (column) and RM (row). Note that the numerical variable RM (the average number of rooms per dwelling in the neighborhood) is grouped into bins of 3-4, 5-6, and so on. Note also the empty cells, denoting that there are no neighborhoods in the dataset with those combinations (e.g., bounding the river and having on average three or four rooms). There are many more possibilities and options for using Excel's pivot tables. We leave it to the reader to explore these using Excel's documentation.



Count of MEDV	
CHAS	Total
0	471
1	35
Grand Total	506

Average of MEDV	CHAS		
RM	0	1	Grand Total
3-4	25.3		25.3
4-5	16.023077		16.02307692
5-6	17.133333	22.21818182	17.48734177
6-7	21.76917	25.91875	22.01598513
7-8	35.964444	44.06666667	36.91764706
8-9	45.7	35.95	44.2
Grand Total	22.093843	28.44	22.53280632

FIGURE 4.4: PIVOT TABLES FOR THE BOSTON HOUSING DATA

In classification tasks, where the goal is to find predictor variables that do a good job of distinguishing between two classes, a good exploratory step is to produce summaries for each class. This can assist in detecting useful predictors that display some separation between the two classes. Data summaries are useful for almost any data mining task and are therefore an important preliminary step for cleaning and understanding the data before carrying out further analyses.

4.4 Correlation Analysis

In datasets with a large number of variables (which are likely to serve as predictors), there is usually much overlap in the information covered by the set of variables. One simple way to find redundancies is to look at a correlation matrix. This shows all the pairwise correlations between variables. Pairs that have a very strong (positive or negative) correlation contain a lot of overlap in information and are good candidates for data reduction by removing one of the variables. Removing variables that are strongly correlated to others is useful for avoiding multicollinearity problems that can arise in various models. (*Multicollinearity* is the presence of two or more predictors sharing the same linear relationship with the outcome variable.)

Correlation analysis is also a good method for detecting duplications of variables in the data. Sometimes, the same variable appears accidentally more than once in the dataset (under a different name) because the dataset was merged from multiple sources, the same phenomenon is measured in different units, and so on. Using correlation table heatmaps, as shown in Chapter 3, can make the task of identifying strong correlations easier.

4.5 Reducing the Number of Categories in Categorical Variables

When a categorical variable has many categories, and this variable is destined to be a predictor, many data mining methods will require converting it into many dummy variables. In particular, a variable with *m* categories will be transformed into *m* — 1 dummy variables. This means that even if we have very few original categorical variables, they can greatly inflate the dimension of the dataset. One way to handle this is to reduce the number of categories by combining close or similar categories. To combine categories requires incorporating expert knowledge and common sense. Pivot tables are useful for this task: We can examine the sizes of the various categories and how the response behaves at each category. Generally, categories that contain very few observations are good candidates for combining with other categories. Use only the categories that are most relevant to the analysis, and label the rest as "other." In classification tasks (with a categorical output), a pivot table broken down by the output classes can help identify categories that do not separate the classes. Those categories too are candidates for inclusion in the "other" category. An example is shown in Figure 4.5, where the distribution of output variable CAT.MEDV is broken down by ZN (treated here as a categorical variable). We can see that the distribution of CAT.MEDV is identical for ZN=17.5, 90, 95, and 100 (where all neighborhoods have CAT.MEDV=1). These four categories can then be combined into a single category. Similarly categories ZN=12.5, 25, 28, 30, and 70 can be combined. Further combination is also possible based on similar bars.

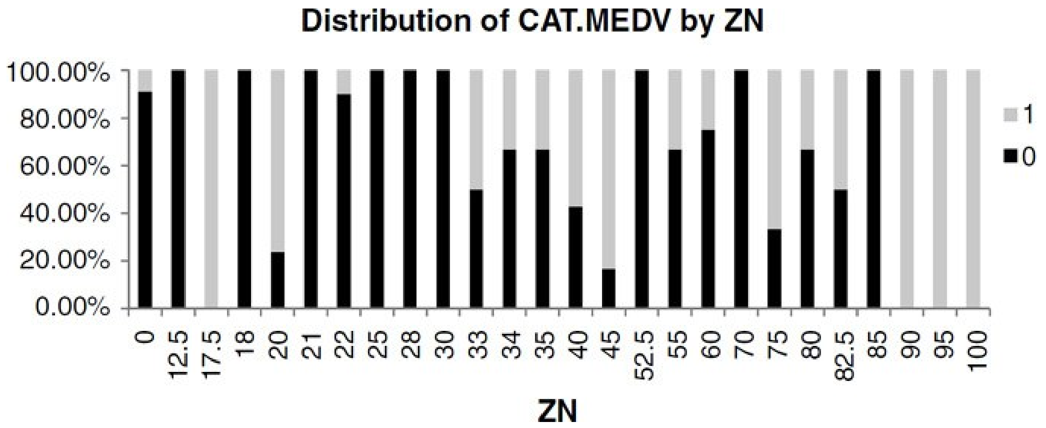


FIGURE 4.5: DISTRIBUTION OF CAT.MEDV (BLACK DENOTES CAT.MEDV=O) BY ZN. SIMILAR BARS INDICATE LOW SEPARATION BETWEEN CLASSES AND CAN BE COMBINED

In a time series context where we might have a categorical variable denoting season (such as month, or hour of day) that will serve as a predictor, reducing categories can be done by examining the time series plot and identifying similar periods. For example, the time plot in Figure 4.6 shows the quarterly revenues of Toys "R" Us between 1992 and 1995. Only quarter 4 periods appear different, and therefore we can combine quarters 1-3 into a single category.

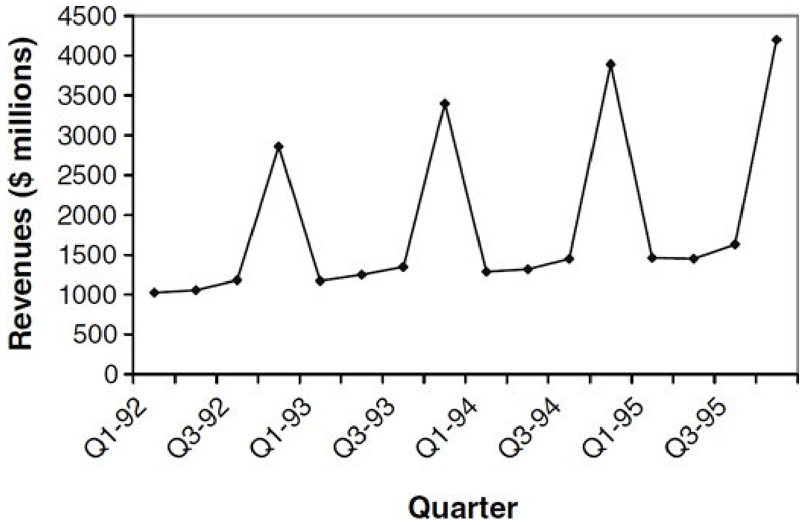


FIGURE 4.6: QUARTERLY REVENUES OF TOYS "R" US, 1992-1995

4.6 Converting a Categorical Variable to a Numerical Variable

Sometimes the categories in a categorical variable represent intervals. Common examples are age group or income group. If the interval values are known (e.g., category 2 is the age interval 20-30), we can replace the categorical value ("2" in the example) with the midinterval value (here "25"). The result will be a numerical variable that no longer requires multiple dummy variables.

4.7 Principal Components Analysis

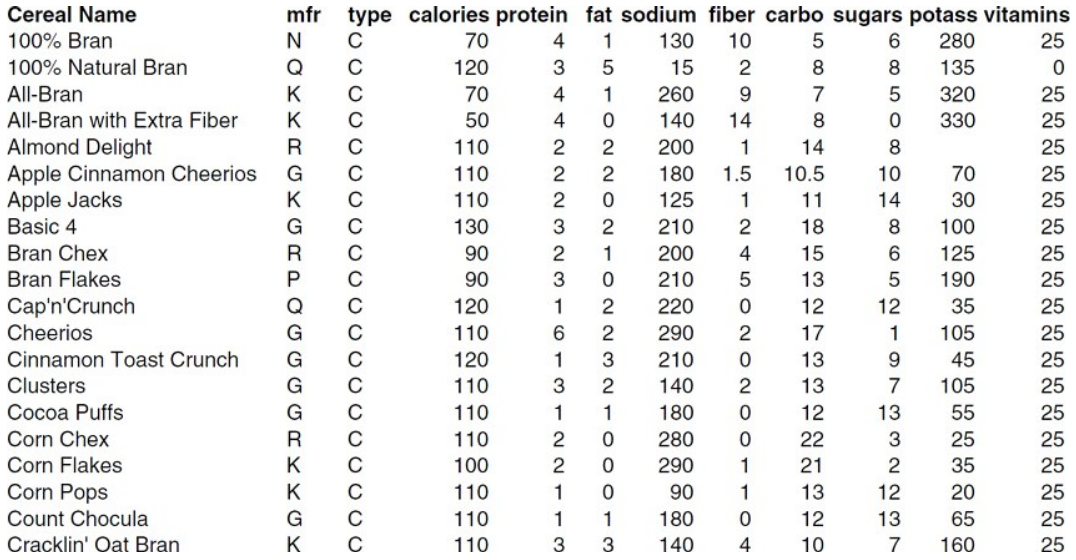
Principal components analysis (PCA) is a useful procedure for reducing the number of predictors in the model by analyzing the input variables. It is especially valuable when we have subsets of measurements that are highly correlated. In that case it provides a few variables (often as few as three) that are weighted linear combinations of the original variables that retain the explanatory power of the full original set. PCA is intended for use with quantitative variables. For categorical variables, other methods, such as correspondence analysis, are more suitable.

Example 2: Breakfast Cereals

Data were collected on the nutritional information and consumer rating of 77 breakfast cereals.[1] For each cereal the data include 13 numerical variables, and we are interested in reducing this dimension. For each cereal the information is based on a bowl of cereal rather than a serving size because most people simply fill a cereal bowl (resulting in constant volume, but not weight). A snapshot of these data is given in Figure 4.7, and the description of the different variables is given in Table 4.2.

Table 4.2: DESCRIPTION OF THE VARIABLES IN THE BREAKFAST CEREAL DATASET

Variable	Description
mfr	Manufacturer of cereal (American Home Food Products, General Mills, Kellogg, etc.)
type	Cold or hot
calories	Calories per serving
protein	Grams of protein
fat	Grams of fat
sodium	Milligrams of sodium
fiber	Grams of dietary fiber
carbo	Grams of complex carbohydrates
sugars	Grams of sugars
potass	Milligrams of potassium
vitamins	Vitamins and minerals: 0, 25, or 100,
	indicating the typical percentage of FDA recommended
shelf	Display shelf (1, 2, or 3, counting from the floor)
weight	Weight in ounces of one serving
cups	Number of cups in one serving
rating	Rating of the cereal calculated by *Consumer Reports*



Cereal Name	mfr	type	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins
100% Bran	N	C	70	4	1	130	10	5	6	280	25
100% Natural Bran	Q	C	120	3	5	15	2	8	8	135	0
All-Bran	K	C	70	4	1	260	9	7	5	320	25
All-Bran with Extra Fiber	K	C	50	4	0	140	14	8	0	330	25
Almond Delight	R	C	110	2	2	200	1	14	8		25
Apple Cinnamon Cheerios	G	C	110	2	2	180	1.5	10.5	10	70	25
Apple Jacks	K	C	110	2	0	125	1	11	14	30	25
Basic 4	G	C	130	3	2	210	2	18	8	100	25
Bran Chex	R	C	90	2	1	200	4	15	6	125	25
Bran Flakes	P	C	90	3	0	210	5	13	5	190	25
Cap'n'Crunch	Q	C	120	1	2	220	0	12	12	35	25
Cheerios	G	C	110	6	2	290	2	17	1	105	25
Cinnamon Toast Crunch	G	C	120	1	3	210	0	13	9	45	25
Clusters	G	C	110	3	2	140	2	13	7	105	25
Cocoa Puffs	G	C	110	1	1	180	0	12	13	55	25
Corn Chex	R	C	110	2	0	280	0	22	3	25	25
Corn Flakes	K	C	100	2	0	290	1	21	2	35	25
Corn Pops	K	C	110	1	0	90	1	13	12	20	25
Count Chocula	G	C	110	1	1	180	0	12	13	65	25
Cracklin' Oat Bran	K	C	110	3	3	140	4	10	7	160	25

FIGURE 4.7: SAMPLE FROM THE 77 BREAKFAST CEREALS DATASET

We focus first on two variables: *calories* and *consumer rating*. These are given in Table 4.3. The average calories across the 75 cereals is 106.88 and the average consumer rating is 42.67. The estimated covariance matrix between the two variables is

$$S = \begin{bmatrix} 379.63 & -188.68 \\ -188.68 & 197.32 \end{bmatrix}.$$

Table 4.3: CEREAL CALORIES AND RATINGS

Cereal	Calories	Rating	Cereal	Calories	Rating
100% Bran	70	68.40297	Just Right Crunchy Nuggets	110	36.52368
100% Natural Bran	120	33.98368			
All-Bran	70	59.42551	Just Right Fruit & Nut	140	36.471512
All-Bran with Extra Fiber	50	93.70491	Kix	110	39.241114
			Life	100	45.328074
Almond Delight	110	34.38484	Lucky Charms	110	26.734515
Apple Cinnamon Cheerios	110	29.50954	Maypo	100	54.850917
			Muesli Raisins, Dates & Almonds	150	37.136863
Apple Jacks	110	33.17409			
Basic 4	130	37.03856	Muesli Raisins, Peaches & Pecans	150	34.139765
Bran Chex	90	49.12025			
Bran Flakes	90	53.31381	Mueslix Crispy Blend	160	30.313351
Cap'n Crunch	120	18.04285	Multi-Grain Cheerios	100	40.105965
Cheerios	110	50.765	Nut&Honey Crunch	120	29.924285
Cinnamon Toast Crunch	120	19.82357	Nutri-Grain Almond-Raisin	140	40.69232
Clusters	110	40.40021	Nutri-grain Wheat	90	59.642837
Cocoa Puffs	110	22.73645	Oatmeal Raisin Crisp	130	30.450843
Corn Chex	110	41.44502	Post Nat. Raisin Bran	120	37.840594
Corn Flakes	100	45.86332	Product 19	100	41.50354
Corn Pops	110	35.78279	Puffed Rice	50	60.756112
Count Chocula	110	22.39651	Puffed Wheat	50	63.005645
Cracklin' Oat Bran	110	40.44877	Quaker Oat Squares	100	49.511874
Cream of Wheat (Quick)	100	64.53382	Quaker Oatmeal	100	50.828392
			Raisin Bran	120	39.259197
Crispix	110	46.89564	Raisin Nut Bran	100	39.7034
Crispy Wheat & Raisins	100	36.1762	Raisin Squares	90	55.333142
			Rice Chex	110	41.998933
Double Chex	100	44.33086	Rice Krispies	110	40.560159
Froot Loops	110	32.20758	Shredded Wheat	80	68.235885
Frosted Flakes	110	31.43597	Shredded Wheat ‘n’ Bran	90	74.472949
Frosted Mini-Wheats	100	58.34514			
			Shredded Wheat spoon size	90	72.801787
Fruit & Fibre Dates, Walnuts & Oats	120	40.91705			
			Smacks	110	31.230054
Fruitful Bran	120	41.01549	Special K	110	53.131324
Fruity Pebbles	110	28.02577	Strawberry Fruit Wheats	90	59.363993
Golden Crisp	100	35.25244			
Golden Grahams	110	23.80404	Total Corn Flakes	110	38.839746
Grape Nuts Flakes	100	52.0769	Total Raisin Bran	140	28.592785
Grape-Nuts	110	53.37101	Total Whole Grain	100	46.658844
Great Grains Pecan	120	45.81172	Triples	110	39.106174
Honey Graham Ohs	120	21.87129	Trix	110	27.753301
Honey Nut Cheerios	110	31.07222	Wheat Chex	100	49.787445

Honey-comb	110	28.74241	Wheaties	100	51.592193
			Wheaties Honey Gold	110	36.187559

It can be seen that the two variables are strongly correlated with a negative correlation of

$$-0.69 = \frac{-188.68}{\sqrt{(379.63)(197.32)}}.$$

Roughly speaking, 69% of the total variation in both variables is actually "covariation," or variation in one variable that is duplicated by similar variation in the other variable. Can we use this fact to reduce the number of variables, while making maximum use of their unique contributions to the overall variation? Since there is redundancy in the information that the two variables contain, it might be possible to reduce the two variables to a single variable without losing too much information. The idea in PCA is to find a linear combination of the two variables that contains most, even if not all, of the information, so that this new variable can replace the two original variables. Information here is in the sense of variability: What can explain the most variability *among* the 77 cereals? The total variability here is the sum of the variances of the two variables, which in this case is 379.63 + 197.32 = 577. This means that *calories* accounts for 66% = 379.63/577 of the total variability, and *rating* for the remaining 34%. If we drop one of the variables for the sake of dimension reduction, we lose at least 34% of the total variability. Can we redistribute the total variability between two new variables in a more polarized way? If so, it might be possible to keep only the one new variable that (hopefully) accounts for a large portion of the total variation.

Figure 4.8 shows a scatterplot of *rating* versus *calories*. The line z_1 is the direction in which the variability of the points is largest. It is the line that captures the most variation in the data ifwe decide to reduce the dimensionality of the data from two to one. Among all possible lines, it is the line for which, if we project the points in the dataset orthogonally to get a set of77 (one-dimensional) values, the variance of the z_1 values will be maximum. This is called the *first principal component*. It is also the line that minimizes the sum-of-squared perpendicular distances from the line. The z_2 axis is chosen to be perpendicular to the z_1-axis. In the case of two variables, there is only one line that is perpendicular to z_1, and it has the second largest variability, but its information is uncorrelated with z_1. This is called the *second principal component*. In general, when we have more than two variables, once we find the direction z_1 with the largest variability, we search among all the orthogonal directions to z_1 for the one with the next highest variability. That is z_2. The idea is then to find the coordinates of these lines and to see how they redistribute the variability.

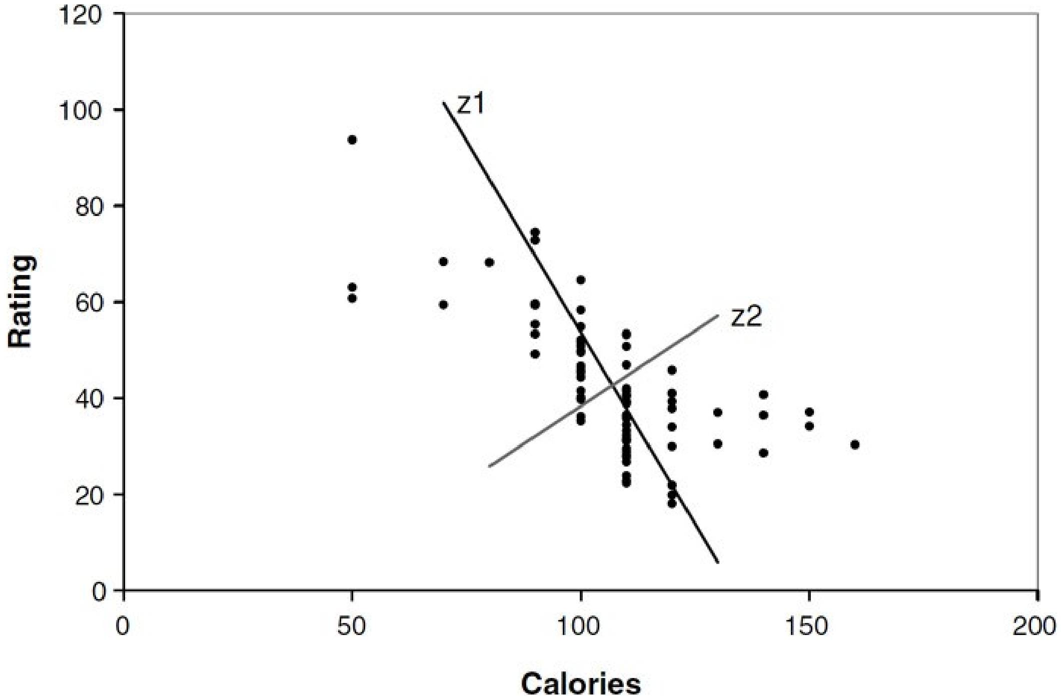
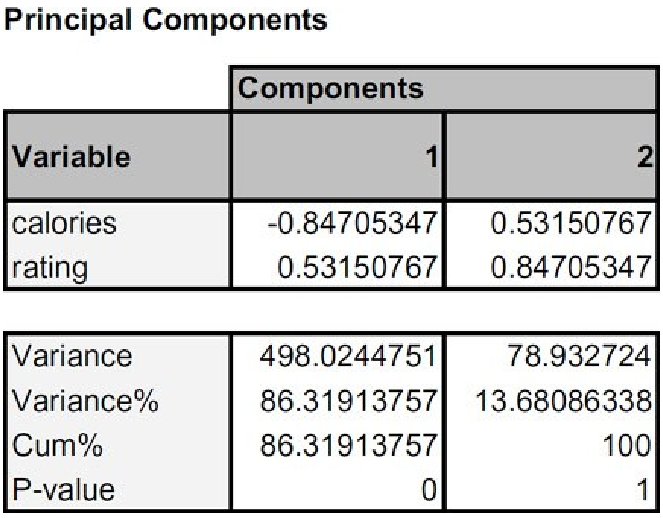


FIGURE 4.8: SCATTERPLOT OF *RATING* VS. *CALORIES* FOR 77 BREAKFAST CEREALS, WITH THE TWO

PRINCIPAL COMPONENT DIRECTIONS



Principal Components

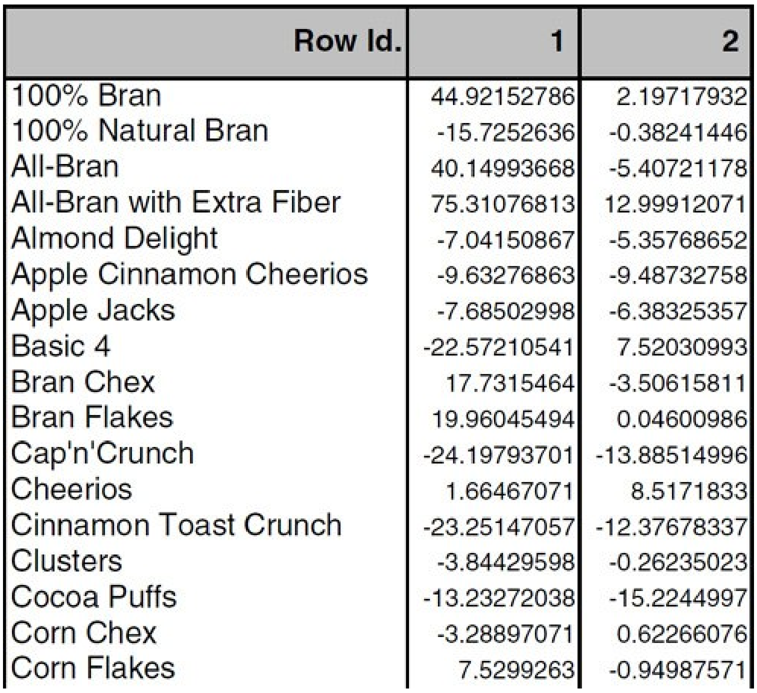
	Components	
Variable	**1**	**2**
calories	-0.84705347	0.53150767
rating	0.53150767	0.84705347

Variance	498.0244751	78.932724
Variance%	86.31913757	13.68086338
Cum%	86.31913757	100
P-value	0	1

FIGURE 4.9: OUTPUT FROM PRINCIPAL COMPONENTS ANALYSIS OF *CALORIES* AND *RATING*

Figure 4.9 shows the XLMiner output from running PCA on these two variables. The principal components table gives the weights that are used to project the original points onto the two new directions. The weights for z_1 are givenby(-0.847, 0.532), and for z_2 they are given by (0.532, 0.847). Figure 4.9 gives the reallocated variance: z_1 accounts for 86% of the total variability and z_2 for the remaining 14%. Therefore, if we drop z_2, we still maintain 86% of the total variability.

The weights are used to compute principal component scores, which are the projected values of *calories* and *rating* onto the new axes (after subtracting the means). Figure 4.10 shows the scores for the two dimensions. The first column is the projection onto z_1 using the weights (−0.847, 0.532). The second column 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.



Row Id.	1	2
100% Bran	44.92152786	2.19717932
100% Natural Bran	-15.7252636	-0.38241446
All-Bran	40.14993668	-5.40721178
All-Bran with Extra Fiber	75.31076813	12.99912071
Almond Delight	-7.04150867	-5.35768652
Apple Cinnamon Cheerios	-9.63276863	-9.48732758
Apple Jacks	-7.68502998	-6.38325357
Basic 4	-22.57210541	7.52030993
Bran Chex	17.7315464	-3.50615811
Bran Flakes	19.96045494	0.04600986
Cap'n'Crunch	-24.19793701	-13.88514996
Cheerios	1.66467071	8.5171833
Cinnamon Toast Crunch	-23.25147057	-12.37678337
Clusters	-3.84429598	-0.26235023
Cocoa Puffs	-13.23272038	-15.2244997
Corn Chex	-3.28897071	0.62266076
Corn Flakes	7.5299263	-0.94987571

FIGURE 4.10: PRINCIPAL SCORES FROM PRINCIPAL COMPONENTS ANALYSIS OF *CALORIES* AND *RATING* FOR THE FIRST 17 CEREALS

Note that the means of the new variables z_1 and z_2 are zero (because we have subtracted the mean of each variable). The sum of the variances var(z_1) + var(z_2) is equal to the sum of the variances of the original variables, *calories* and *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

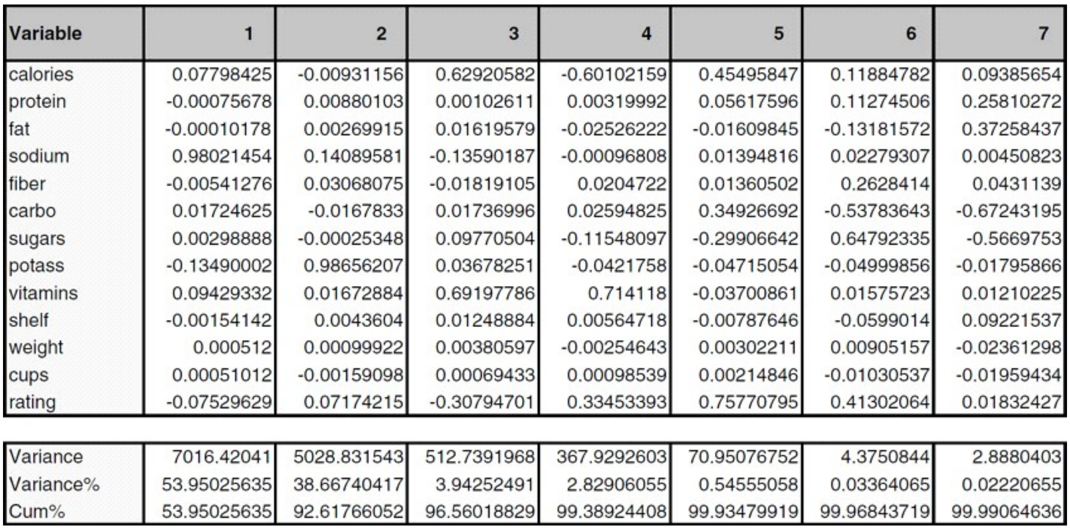
Let us formalize the procedure described above so that it can easily be generalized to *p*> > 2 variables. Denote by X_1, X_2, …, X_p the original *p* variables. InPCAwe are looking for a set of new variables Z_1, Z_2, …, Z_p that are weighted averages of the original variables (after subtracting their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

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

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.

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 4.11. For simplicity, we removed three cereals that contained missing values. 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.



Variable	1	2	3	4	5	6	7
calories	0.07798425	-0.00931156	0.62920582	-0.60102159	0.45495847	0.11884782	0.09385654
protein	-0.00075678	0.00880103	0.00102611	0.00319992	0.05617596	0.11274506	0.25810272
fat	-0.00010178	0.00269915	0.01619579	-0.02526222	-0.01609845	-0.13181572	0.37258437
sodium	0.98021454	0.14089581	-0.13590187	-0.00096808	0.01394816	0.02279307	0.00450823
fiber	-0.00541276	0.03068075	-0.01819105	0.0204722	0.01360502	0.2628414	0.0431139
carbo	0.01724625	-0.0167833	0.01736996	0.02594825	0.34926692	-0.53783643	-0.67243195
sugars	0.00298888	-0.00025348	0.09770504	-0.11548097	-0.29906642	0.64792335	-0.5669753
potass	-0.13490002	0.98656207	0.03678251	-0.0421758	-0.04715054	-0.04999856	-0.01795866
vitamins	0.09429332	0.01672884	0.69197786	0.714118	-0.03700861	0.01575723	0.01210225
shelf	-0.00154142	0.0043604	0.01248884	0.00564718	-0.00787646	-0.0599014	0.09221537
weight	0.000512	0.00099922	0.00380597	-0.00254643	0.00302211	0.00905157	-0.02361298
cups	0.00051012	-0.00159098	0.00069433	0.00098539	0.00214846	-0.01030537	-0.01959434
rating	-0.07529629	0.07174215	-0.30794701	0.33453393	0.75770795	0.41302064	0.01832427

Variance	7016.42041	5028.831543	512.7391968	367.9292603	70.95076752	4.3750844	2.8880403
Variance%	53.95025635	38.66740417	3.94252491	2.82906055	0.54555058	0.03364065	0.02220655
Cum%	53.95025635	92.61766052	96.56018829	99.38924408	99.93479919	99.96843719	99.99064636

FIGURE 4.11: PCA OUTPUT USING ALL 13 NUMERICAL VARIABLES IN THE BREAKFAST CEREALS DATASET. RESULTS ARE GIVEN FOR THE FIRST SEVEN PRINCIPAL COMPONENTS

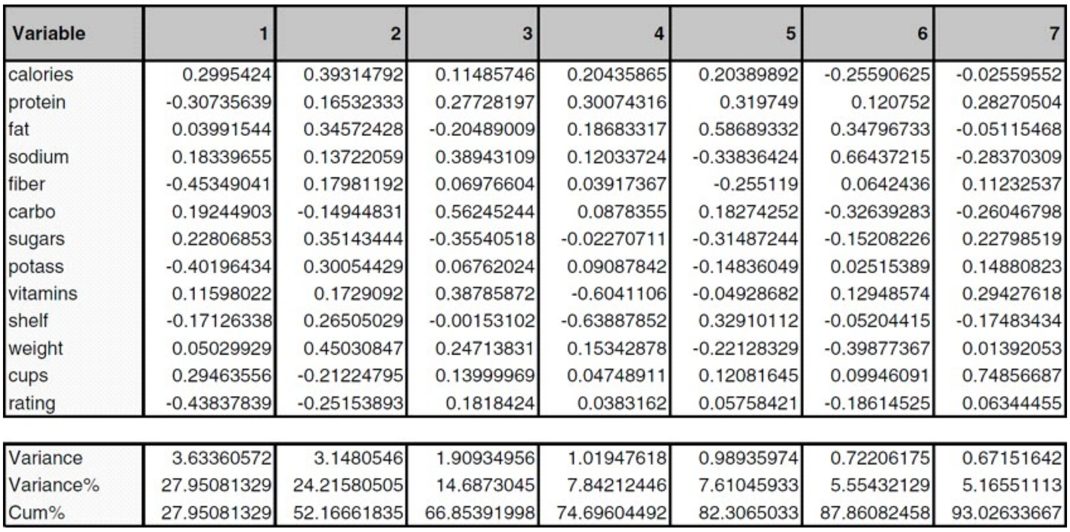
Normalizing the Data

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 measuring how much sodium is in the cereal. Similarly, the second principal component seems to be measuring the amount of potassium. Since both these 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), 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 quite differing 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 that 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 4.12. Now we find that we need 7 principal components to account for 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 positive weights) versus (2) protein, fiber, potassium, and consumer rating (large negative weights). High scores on principal component 1 mean that the cereal is high in calories and the amount per bowl, and low in protein, fiber, 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.



Variable	1	2	3	4	5	6	7
calories	0.2995424	0.39314792	0.11485746	0.20435865	0.20389892	-0.25590625	-0.02559552
protein	-0.30735639	0.16532333	0.27728197	0.30074316	0.319749	0.120752	0.28270504
fat	0.03991544	0.34572428	-0.20489009	0.18683317	0.58689332	0.34796733	-0.05115468
sodium	0.18339655	0.13722059	0.38943109	0.12033724	-0.33836424	0.66437215	-0.28370309
fiber	-0.45349041	0.17981192	0.06976604	0.03917367	-0.255119	0.0642436	0.11232537
carbo	0.19244903	-0.14944831	0.56245244	0.0878355	0.18274252	-0.32639283	-0.26046798
sugars	0.22806853	0.35143444	-0.35540518	-0.02270711	-0.31487244	-0.15208226	0.22798519
potass	-0.40196434	0.30054429	0.06762024	0.09087842	-0.14836049	0.02515389	0.14880823
vitamins	0.11598022	0.1729092	0.38785872	-0.6041106	-0.04928682	0.12948574	0.29427618
shelf	-0.17126338	0.26505029	-0.00153102	-0.63887852	0.32910112	-0.05204415	-0.17483434
weight	0.05029929	0.45030847	0.24713831	0.15342878	-0.22128329	-0.39877367	0.01392053
cups	0.29463556	-0.21224795	0.13999969	0.04748911	0.12081645	0.09946091	0.74856687
rating	-0.43837839	-0.25153893	0.1818424	0.0383162	0.05758421	-0.18614525	0.06344455

Variance	3.63360572	3.1480546	1.90934956	1.01947618	0.98935974	0.72206175	0.67151642
Variance%	27.95081329	24.21580505	14.6873045	7.84212446	7.61045933	5.55432129	5.16551113
Cum%	27.95081329	52.16661835	66.85391998	74.69604492	82.3065033	87.86082458	93.02633667

FIGURE 4.12: PCA OUTPUT USING ALL *NORMALIZED* 13 NUMERICAL VARIABLES IN THE BREAKFAST CEREALS DATASET. RESULTS ARE GIVEN FOR THE FIRST SEVEN PRINCIPAL COMPONENTS

When the data can be reduced to two dimensions, a useful plot is a scatterplot of the first versus the second principal scores with labels for the observations (if the dataset is not too large). To illustrate this, Figure 4.13 displays the first two principal component scores for the breakfast cereals.

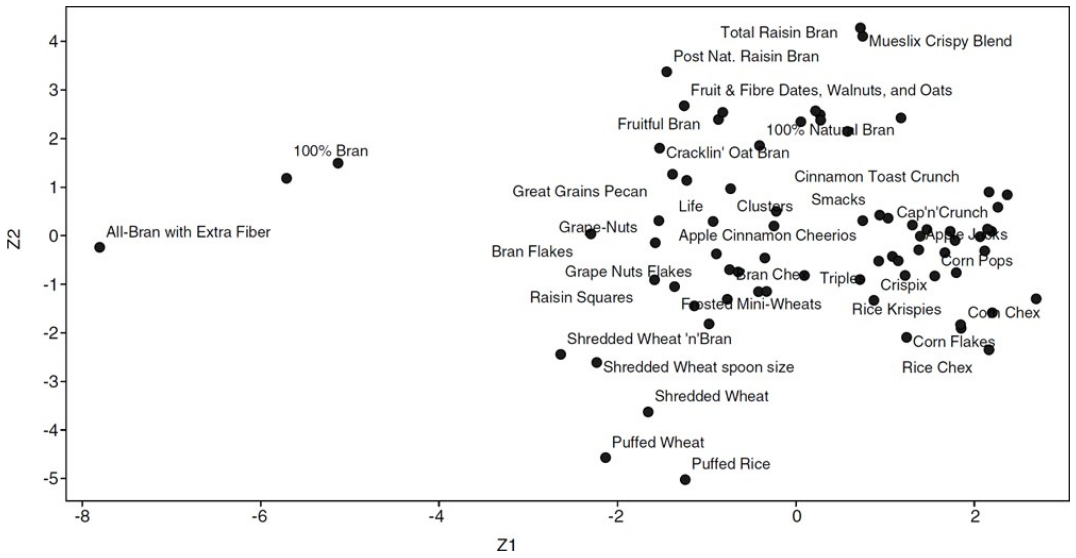


FIGURE 4.13: SCATTERPLOT OF THE SECOND VS. FIRST PRINCIPAL COMPONENTS SCORES FOR THE NORMALIZED BREAKFAST CEREAL OUTPUT

We can see that as we move from left (bran cereals) to right, 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 observations can be found. For instance, we see here that children's cereals are close together on the middle-right part of the plot.

Using Principal Components for Classification and Prediction

When the goal of the data reduction is to have a smaller set of variables that will serve as predictors, we can proceed as follows: Apply PCA to the training data. Use the output to determine the number of principal components to be retained. The predictors in the model now use the (reduced number of) principal scores columns. For the validation set we can use the weights computed from the training data to obtain a set of principal scores by applying the weights to the variables in the validation set. These new variables are then treated as the predictors.

[1] The data are available at http://lib.stat.cmu.edu/DASL/Stories/HealthyBreakfast. html.

4.8 Dimension Reduction Using Regression Models

In this chapter we discussed methods for reducing the number of columns using summary statistics, plots, and principal components analysis. All these are considered exploratory methods. Some of them completely ignore the output variable (e.g., PCA), whereas in other methods we informally try to incorporate the relationship between the predictors and the output variable (e.g., combining similar categories, in terms of their behavior with y). Another approach to reducing the number of predictors, which directly considers the predictive or classification task, is by fitting a regression model. For prediction a linear regression model is used (see Chapter 6), and for classification a logistic regression model (see Chapter 10) is used. In both cases we can employ subset selection procedures that algorithmically choose a subset of variables among the larger set (see details in the relevant chapters).

Fitted regression models can also be used to further combine similar categories: categories that have coefficients that are not statistically significant (i.e., have a high *p*-value) can be combined with the reference category because their distinction from the reference category appears to have no significant effect on the output variable. Moreover, categories that have similar coefficient values (and the same sign) can often be combined because their effect on the output variable is similar. See the example in Chapter 10 on predicting delayed flights for an illustration of how regression models can be used for dimension reduction.

4.9 Dimension Reduction Using Classification and Regression Trees

Another method for reducing the number of columns and for combining categories of a categorical variable is by applying classification and regression trees (see Chapter 9). Classification trees are used for classification tasks and regression trees for prediction tasks. In both cases the algorithm creates binary splits on the predictors that best classify/predict the outcome (e.g., above/below age 30). Although we defer the detailed discussion to Chapter 9, we note here that the resulting tree diagram can be used for determining the important predictors. Predictors (numerical or categorical) that do not appear in the tree can be removed. Similarly, categories that do not appear in the tree can be combined.

Problems

4.1 **Breakfast Cereals**. Use the data for the breakfast cereal example in Section 4.7 to explore and summarize the data as follows: (Note that a few records contain missing values; since there are just a few, a simple solution is to remove them first. You can use the "Missing Data Handling" utility in XLMiner.)

a. Which variables are quantitative/numerical? Which are ordinal? Which are nominal?

b. Create a table with the average, median, min, max, and standard deviation for each of the quantitative variables. This can be done through Excel's functions or Excel's *Tools* > *DataAnalysis* > *DescriptiveStatistics* menu.

c. Use XLMiner to plot a histogram for each of the quantitative variables. Based on the histograms and summary statistics, answer the following questions:

   i. Which variables have the largest variability?

   ii. Which variables seem skewed?

   iii. Are there any values that seem extreme?

d. Use XLMiner to plot a side-by-side boxplot comparing the calories in hot versus cold cereals. What does this plot show us?

e. Use XLMiner to plot a side-by-side boxplot of consumer rating as a function of the shelf height. If we were to predict consumer rating from shelf height, does it appear that we need to keep all three categories of shelf height?

f. Compute the correlation table for the quantitative variable (use Excel's *Tools* > *Data-Analysis* > *Correlation* menu). In addition, use XLMiner to generate a matrix plot for these variables.

   i. Which pair of variables is most strongly correlated?

   ii. How can we reduce the number of variables based on these correlations?

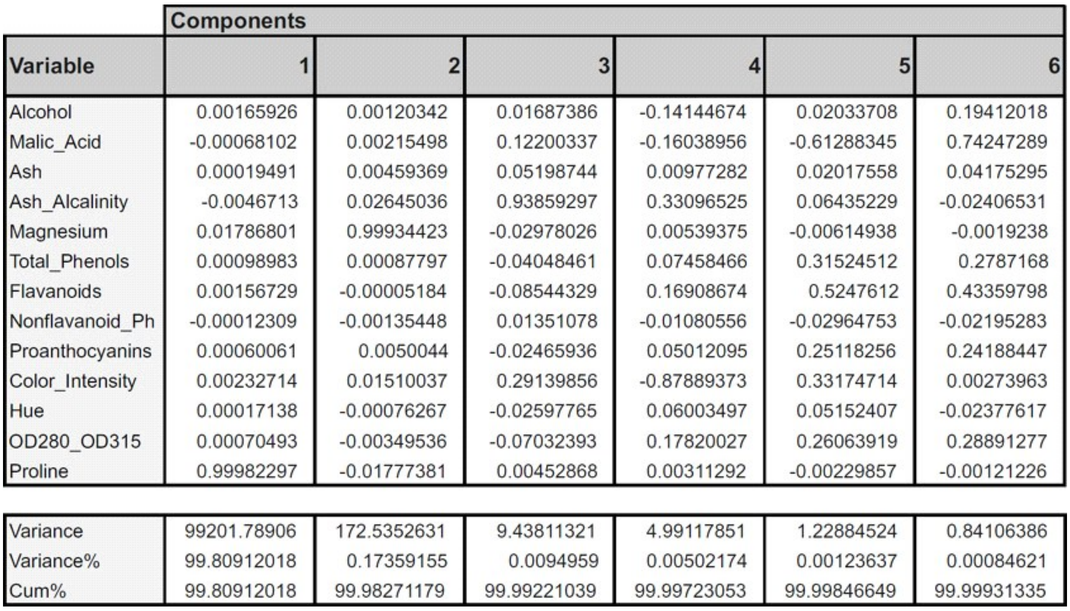
   iii. How would the correlations change if we normalized the data first?

g. Consider the first column on the left in Figure 4.11. Describe briefly what this column represents.

4.2 **Chemical Features of Wine**. Figure 4.14 shows the PCA output on data (nonnormal-ized) in which the variables represent chemical characteristics of wine, and each case is a different wine.

a. The data are in the file Wine.xls. Consider the row near the bottom labeled "Variance" Explain why column 1's variance is so much greater than that of any other column.

b. Comment on the use of normalization (standardization) in part (a).



	Components					
Variable	1	2	3	4	5	6
Alcohol	0.00165926	0.00120342	0.01687386	-0.14144674	0.02033708	0.19412018
Malic_Acid	-0.00068102	0.00215498	0.12200337	-0.16038956	-0.61288345	0.74247289
Ash	0.00019491	0.00459369	0.05198744	0.00977282	0.02017558	0.04175295
Ash_Alcalinity	-0.0046713	0.02645036	0.93859297	0.33096525	0.06435229	-0.02406531
Magnesium	0.01786801	0.99934423	-0.02978026	0.00539375	-0.00614938	-0.0019238
Total_Phenols	0.00098983	0.00087797	-0.04048461	0.07458466	0.31524512	0.2787168
Flavanoids	0.00156729	-0.00005184	-0.08544329	0.16908674	0.5247612	0.43359798
Nonflavanoid_Ph	-0.00012309	-0.00135448	0.01351078	-0.01080556	-0.02964753	-0.02195283
Proanthocyanins	0.00060061	0.0050044	-0.02465936	0.05012095	0.25118256	0.24188447
Color_Intensity	0.00232714	0.01510037	0.29139856	-0.87889373	0.33174714	0.00273963
Hue	0.00017138	-0.00076267	-0.02597765	0.06003497	0.05152407	-0.02377617
OD280_OD315	0.00070493	-0.00349536	-0.07032393	0.17820027	0.26063919	0.28891277
Proline	0.99982297	-0.01777381	0.00452868	0.00311292	-0.00229857	-0.00121226

Variance	99201.78906	172.5352631	9.43811321	4.99117851	1.22884524	0.84106386
Variance%	99.80912018	0.17359155	0.0094959	0.00502174	0.00123637	0.00084621
Cum%	99.80912018	99.98271179	99.99221039	99.99723053	99.99846649	99.99931335

FIGURE 4.14: PRINCIPAL COMPONENTS OF NONNORMALIZED WINE DATA

4.3 **University Rankings**. The dataset on American college and university rankings (available from www.dataminingbook.com) contains information on 1302 American colleges and universities offering an undergraduate program. For each university there are 17 measurements that include continuous measurements (such as tuition and graduation rate) and categorical measurements (such as location by state and whether it is a private or a public school).

a. Remove all categorical variables. Then remove all records with missing numerical measurements from the dataset (by creating a new worksheet).

b. Conduct a principal components analysis on the cleaned data and comment on the results. Should the data be normalized? Discuss what characterizes the components you consider key.

4.4 **Sales of Toyota Corolla Cars**. The file ToyotaCorolla.xls contains data on used cars (Toyota Corollas) on sale during late summer of 2004 in The Netherlands. It has 1436 records containing details on 38 attributes, including Price, Age, Kilometers, HP, and other specifications. The goal will be to predict the price of a used Toyota Corolla based on its specifications.

a. Identify the categorical variables.

b. Explain the relationship between a categorical variable and the series of binary dummy variables derived from it.

c. How many dummy binary variables are required to capture the information in a categorical variable with *N* categories?

d. Using XLMiner's data utilities, convert the categorical variables in this dataset into dummy binaries, and explain in words, for one record, the values in the derived binary dummies.

e. Use Excel's correlation command (*Tools* > *DataAnalysis* > *Correlation* menu) to produce a correlation matrix and XLMiner's matrix plot to obtain a matrix of all scatterplots. Comment on the relationships among variables.