Dimensionality Reduction: Principal Components Analysis

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. The accuracy and reliability of a classification or prediction model will suffer if we include highly correlated variables or variables that are unrelated to the outcome of interest because of over fitting. In model deployment also superfluous variables can increase costs due to 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 is therefore finding ways to reduce dimensionality without sacrificing accuracy.

A useful procedure for this purpose is to analyze the principal components of the input variables. It is especially valuable when we have subsets of measurements that are measured on the same scale and are highly correlated. In that case it provides a few (often less than three) variables that are weighted combinations of the original variables that retain the explanatory power of the full original set.

Example 1: Head Measurements of First Adult Sons

The data below give 25 pairs of head measurements for first adult sons in a sample [1].

First Adult Son	
Head Length	Head Breadth
(x1)	(x1)
191	155
195	149
181	148
183	153
176	144
208	157
189	150
197	159
188	152
192	150
179	158
183	147
174	150
190	159
188	151
163	137
195	155
186	153
181	145
175	140
192	154
174	143
176	139
197	167
190	163

For this data the means of the variables x1 and x2 are 185.7 and 151.1 and the covariance matrix, S = $\begin{bmatrix} 95.29 & 52.87 \\ 52.87 & 54.36 \end{bmatrix}$

Figure 1 below shows the scatter plot of points (x1, x2). The principal component directions are shown by the axes z1 and z2 that are centered at the means of x1 and x2. The line z1 is the direction of the first principal component of the data. It is the line that captures the most variation in the data if we decide to reduce the dimensionality of the data from two to one. Amongst all possible lines it is the line that if we project the points in the data set orthogonally to get a set of 25 (one dimensional) values using the z1 coordinate, the variance of the z1 values will be maximum. It is also the line that minimizes the sum of squared perpendicular distances from the line. (Show why this follows from Pythagoras' theorem. How is this line different from the regression line of x2 on x1?) The z2 axis is perpendicular to the z1 axis.

The directions of the axes are given by the eigenvectors of S. For our example the eigenvalues are 131.52 and 18.14. The eigenvector corresponding to the larger eigenvalue is (0.825,0.565) and gives us the direction of the z1 axis. The eigenvector corresponding to the smaller eigenvalue is (- 0.565, 0.825) and this is the direction of the z2 axis.

The lengths of the major and minor axes of the ellipse that would enclose about 40% of the points if the points had a bivariate normal distribution are the square roots of the eigenvalues. This corresponds to rule for being within one standard deviation of the mean for the (univariate) normal distribution. Similarly in that case doubling the axes lengths of the ellipse will enclose 86% of the points and tripling it would enclose 99% of the points. For our example the length of the major axis is √131.5 = 11.47 and √18.14 = 4.26. In Figure 1 the inner ellipse has these axes lengths while the outer ellipse has axes with twice these lengths.

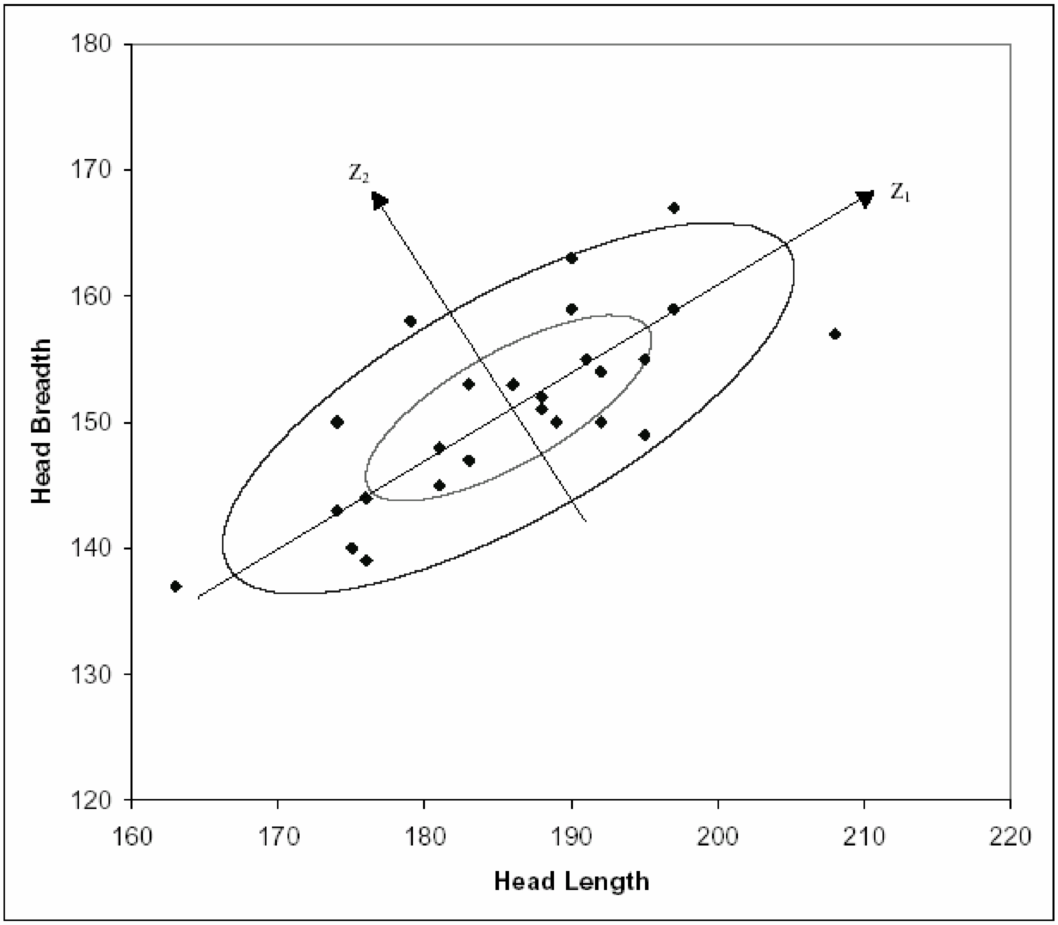


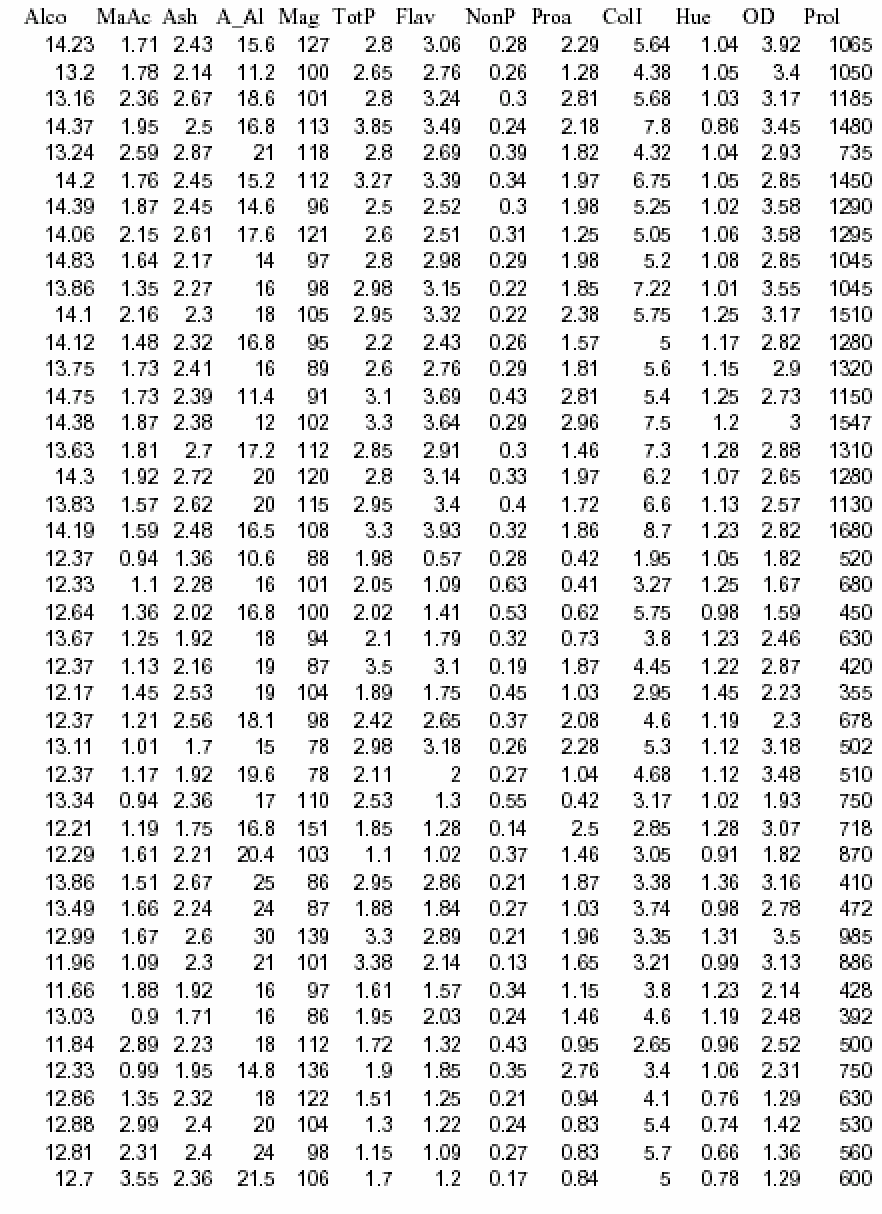
Figure 1

The values of z1 and z2 for the observations are known as the principal component scores and are shown below. The scores are computed as the inner products of the data points and the first and second eigenvectors (in order of decreasing eigenvalue).
The means of z1 and z2 are zero. This follows from our choice of the origin for the (z1, z2) coordinate system to be the means of x1 and x2. The variances are more interesting. The variances of z1 and z2 are 131.5 and 18.14 respectively. The first principal component, z1, accounts for 88% 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.

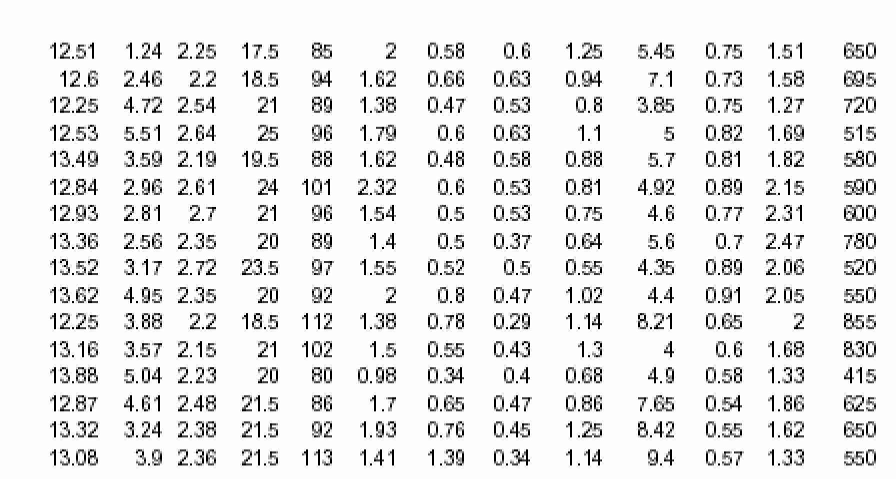
Example 2: Characteristics of Wine

The data in Table 2 gives measurements on 13 characteristics of 60 different wines from a region. Let us see how principal component analysis would enable us to reduce the number of dimensions in the data.

Table 2

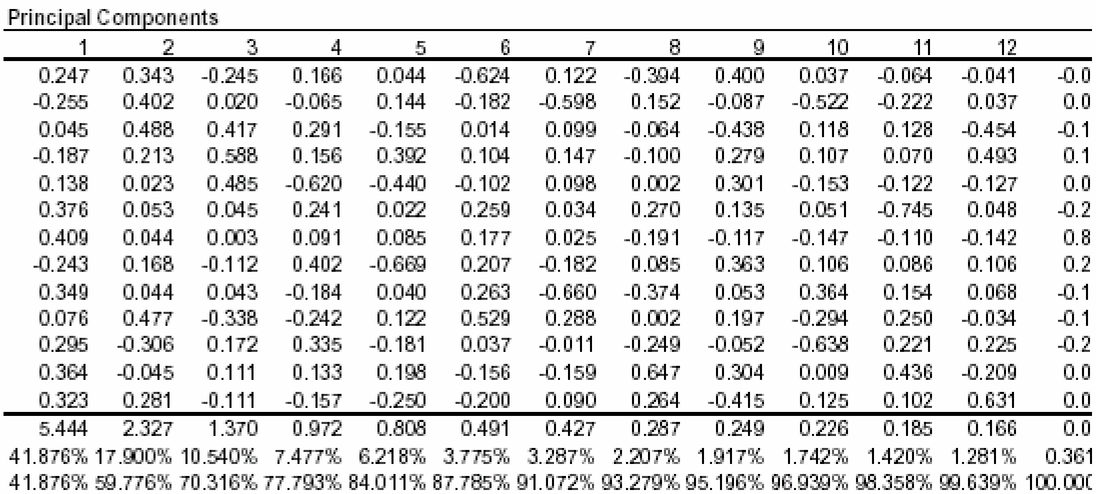


Alco	MaAc	Ash	A_Al	Mag	TotP	Flav	NonP	Proa	ColI	Hue	OD	Prol
14.23	1.71	2.43	15.6	127	2.8	3.06	0.28	2.29	5.64	1.04	3.92	1065
13.2	1.78	2.14	11.2	100	2.65	2.76	0.26	1.28	4.38	1.05	3.4	1050
13.16	2.36	2.67	18.6	101	2.8	3.24	0.3	2.81	5.68	1.03	3.17	1185
14.37	1.95	2.5	16.8	113	3.85	3.49	0.24	2.18	7.8	0.86	3.45	1480
13.24	2.59	2.87	21	118	2.8	2.69	0.39	1.82	4.32	1.04	2.93	735
14.2	1.76	2.45	15.2	112	3.27	3.39	0.34	1.97	6.75	1.05	2.85	1450
14.39	1.87	2.45	14.6	96	2.5	2.52	0.3	1.98	5.25	1.02	3.58	1290
14.06	2.15	2.61	17.6	121	2.6	2.51	0.31	1.25	5.05	1.06	3.58	1295
14.83	1.64	2.17	14	97	2.8	2.98	0.29	1.98	5.2	1.08	2.85	1045
13.86	1.35	2.27	16	98	2.98	3.15	0.22	1.85	7.22	1.01	3.55	1045
14.1	2.16	2.3	18	105	2.95	3.32	0.22	2.38	5.75	1.25	3.17	1510
14.12	1.48	2.32	16.8	95	2.2	2.43	0.26	1.57	5	1.17	2.82	1280
13.75	1.73	2.41	16	89	2.6	2.76	0.29	1.81	5.6	1.15	2.9	1320
14.75	1.73	2.39	11.4	91	3.1	3.69	0.43	2.81	5.4	1.25	2.73	1150
14.38	1.87	2.38	12	102	3.3	3.64	0.29	2.96	7.5	1.2	3	1547
13.63	1.81	2.7	17.2	112	2.85	2.91	0.3	1.46	7.3	1.28	2.88	1310
14.3	1.92	2.72	20	120	2.8	3.14	0.33	1.97	6.2	1.07	2.65	1280
13.83	1.57	2.62	20	115	2.95	3.4	0.4	1.72	6.6	1.13	2.57	1130
14.19	1.59	2.48	16.5	108	3.3	3.93	0.32	1.86	8.7	1.23	2.82	1680
12.37	0.94	1.36	10.6	88	1.98	0.57	0.28	0.42	1.95	1.05	1.82	520
12.33	1.1	2.28	16	101	2.05	1.09	0.63	0.41	3.27	1.25	1.67	680
12.64	1.36	2.02	16.8	100	2.02	1.41	0.53	0.62	5.75	0.98	1.59	450
13.67	1.25	1.92	18	94	2.1	1.79	0.32	0.73	3.8	1.23	2.46	630
12.37	1.13	2.16	19	87	3.5	3.1	0.19	1.87	4.45	1.22	2.87	420
12.17	1.45	2.53	19	104	1.89	1.75	0.45	1.03	2.95	1.45	2.23	355
12.37	1.21	2.56	18.1	98	2.42	2.65	0.37	2.08	4.6	1.19	2.3	678
13.11	1.01	1.7	15	78	2.98	3.18	0.26	2.28	5.3	1.12	3.18	502
12.37	1.17	1.92	19.6	78	2.11	2	0.27	1.04	4.68	1.12	3.48	510
13.34	0.94	2.36	17	110	2.53	1.3	0.55	0.42	3.17	1.02	1.93	750
12.21	1.19	1.75	16.8	151	1.85	1.28	0.14	2.5	2.85	1.28	3.07	718
12.29	1.61	2.21	20.4	103	1.1	1.02	0.37	1.46	3.05	0.91	1.82	870
13.86	1.51	2.67	25	86	2.95	2.86	0.21	1.87	3.38	1.36	3.16	410
13.49	1.66	2.24	24	87	1.88	1.84	0.27	1.03	3.74	0.98	2.78	472
12.99	1.67	2.6	30	139	3.3	2.89	0.21	1.96	3.35	1.31	3.5	985
11.96	1.09	2.3	21	101	3.38	2.14	0.13	1.65	3.21	0.99	3.13	886
11.66	1.88	1.92	16	97	1.61	1.57	0.34	1.15	3.8	1.23	2.14	428
13.03	0.9	1.71	16	86	1.95	2.03	0.24	1.46	4.6	1.19	2.48	392
11.84	2.89	2.23	18	112	1.72	1.32	0.43	0.95	2.65	0.96	2.52	500
12.33	0.99	1.95	14.8	136	1.9	1.85	0.35	2.76	3.4	1.06	2.31	750
12.86	1.35	2.32	18	122	1.51	1.25	0.21	0.94	4.1	0.76	1.29	630
12.88	2.99	2.4	20	104	1.3	1.22	0.24	0.83	5.4	0.74	1.42	530
12.81	2.31	2.4	24	98	1.15	1.09	0.27	0.83	5.7	0.66	1.36	560
12.7	3.55	2.36	21.5	106	1.7	1.2	0.17	0.84	5	0.78	1.29	600



12.51	1.24	2.25	17.5	85	2	0.58	0.6	1.25	5.45	0.75	1.51	650
12.6	2.46	2.2	18.5	94	1.62	0.66	0.63	0.94	7.1	0.73	1.58	695
12.25	4.72	2.54	21	89	1.38	0.47	0.53	0.8	3.85	0.75	1.27	720
12.53	5.51	2.64	25	96	1.79	0.6	0.63	1.1	5	0.82	1.69	515
13.49	3.59	2.19	19.5	88	1.62	0.48	0.58	0.88	5.7	0.81	1.82	580
12.84	2.96	2.61	24	101	2.32	0.6	0.53	0.81	4.92	0.89	2.15	590
12.93	2.81	2.7	21	96	1.54	0.5	0.53	0.75	4.6	0.77	2.31	600
13.36	2.56	2.35	20	89	1.4	0.5	0.37	0.64	5.6	0.7	2.47	780
13.52	3.17	2.72	23.5	97	1.55	0.52	0.5	0.55	4.35	0.89	2.06	520
13.62	4.95	2.35	20	92	2	0.8	0.47	1.02	4.4	0.91	2.05	550
12.25	3.88	2.2	18.5	112	1.38	0.78	0.29	1.14	8.21	0.65	2	855
13.16	3.57	2.15	21	102	1.5	0.55	0.43	1.3	4	0.6	1.68	830
13.88	5.04	2.23	20	80	0.98	0.34	0.4	0.68	4.9	0.58	1.33	415
12.87	4.61	2.48	21.5	86	1.7	0.65	0.47	0.86	7.65	0.54	1.86	625
13.32	3.24	2.38	21.5	92	1.93	0.76	0.45	1.25	8.42	0.55	1.62	650
13.08	3.9	2.36	21.5	113	1.41	1.39	0.34	1.14	9.4	0.57	1.33	550

The output from running a principal components analysis on this data is shown in Output1 below. The rows of Output1 are in the same order as the columns of Table 1 so that for example row 1 for each principal component gives the weight for alchohol and row 13 gives the weight for proline.



Principal Components

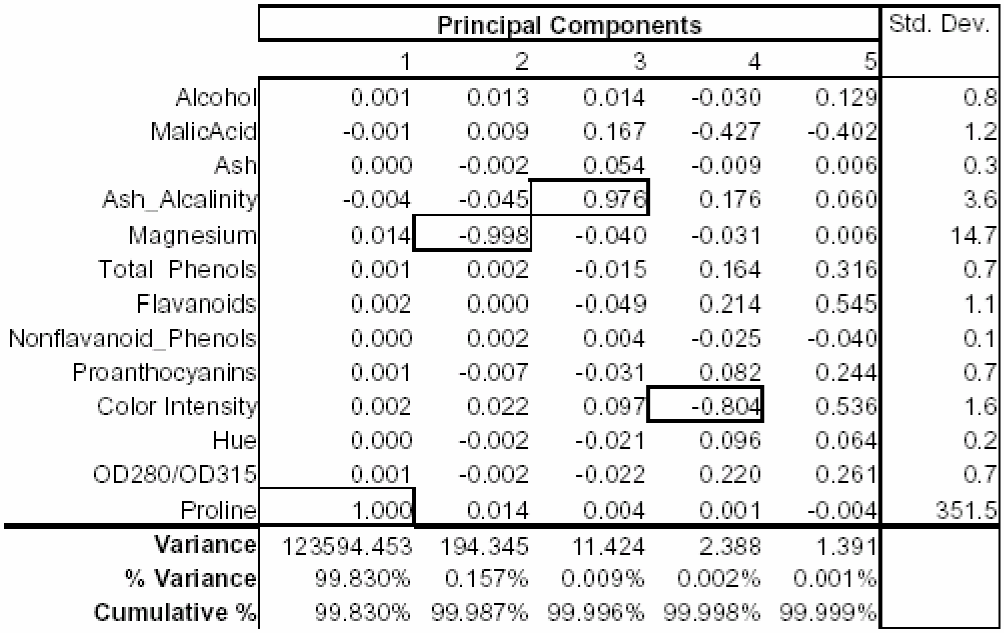
1	2	3	4	5	6	7	8	9	10	11	12	
0.247	0.343	-0.245	0.166	0.044	-0.624	0.122	-0.394	0.400	0.037	-0.064	-0.041	-0.0
-0.255	0.402	0.020	-0.065	0.144	-0.182	-0.598	0.152	-0.087	-0.522	-0.222	0.037	0.0
0.045	0.488	0.417	0.291	-0.155	0.014	0.099	-0.064	-0.438	0.118	0.128	-0.454	-0.1
-0.187	0.213	0.588	0.156	0.392	0.104	0.147	-0.100	0.279	0.107	0.070	0.493	0.1
0.138	0.023	0.485	-0.620	-0.440	-0.102	0.098	0.002	0.301	-0.153	-0.122	-0.127	0.0
0.376	0.053	0.045	0.241	0.022	0.259	0.034	0.270	0.135	0.051	-0.745	0.048	-0.2
0.409	0.044	0.003	0.091	0.085	0.177	0.025	-0.191	-0.117	-0.147	-0.110	-0.142	0.8
-0.243	0.168	-0.112	0.402	-0.669	0.207	-0.182	0.085	0.363	0.106	0.086	0.106	0.2
0.349	0.044	0.043	-0.184	0.040	0.263	-0.660	-0.374	0.053	0.364	0.154	0.068	-0.1
0.076	0.477	-0.338	-0.242	0.122	0.529	0.288	0.002	0.197	-0.294	0.250	-0.034	-0.1
0.295	-0.306	0.172	0.335	-0.181	0.037	-0.011	-0.249	-0.052	-0.638	0.221	0.225	-0.2
0.364	-0.045	0.111	0.133	0.198	-0.156	-0.159	0.647	0.304	0.009	0.436	-0.209	0.0
0.323	0.281	-0.111	-0.157	-0.250	-0.200	0.090	0.264	-0.415	0.125	0.102	0.631	0.0
5.444	2.327	1.370	0.972	0.808	0.491	0.427	0.287	0.249	0.226	0.185	0.166	0.0
41.876%	17.900%	10.540%	7.477%	6.218%	3.775%	3.287%	2.207%	1.917%	1.742%	1.420%	1.281%	0.361
41.876%	59.776%	70.316%	77.793%	84.011%	87.785%	91.072%	93.279%	95.196%	96.939%	98.358%	99.639%	100.00

Notice that the first five components account for more than 80% 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 half the number of original dimensions in the data. A further advantage of the principal components compared to the original data is that it 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.

The principal components shown in Output 1 were computed after after 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 standardization is to give all variables equal importance in terms of the variability. The question of when to standardize has to be answered using information of the nature of the data. When the units of measurement are common for the variables as for example dollars it would generally be desirable not to rescale the data for unit variance. If the variables are measured in quite differing units so that it is unclear how to compare the variability of different variables, it is advisable to scale for unit variance, so that changes in units of measurement do not change the principal component weights. In the rare situations where we can give relative weights to variables we would multiply the unit scaled variables by these weights before doing the principal components analysis.

Example2 (continued)
Rescaling variables in the wine data is a important due to the heterogenous nature of the variables. The first five principal components computed on ther raw unscaled data are shown in Table 3. Notice that the variable Proline is the first principal component and it explains almost all the variance in the data. This is because its standard deviation is 351 compared to the next largest standard deviation of 15 for the variable Magnesium. The second principal component is Magnesium. The standard deviations of all the other variables are about 1% (or less) than that of Proline.

Table 3



	Principal Components					Std. Dev.
	1	2	3	4	5	
Alcohol	0.001	0.013	0.014	-0.030	0.129	0.8
MalicAcid	-0.001	0.009	0.167	-0.427	-0.402	1.2
Ash	0.000	-0.002	0.054	-0.009	0.006	0.3
Ash_Alcalinity	-0.004	-0.045	0.976	0.176	0.060	3.6
Magnesium	0.014	-0.998	-0.040	-0.031	0.006	14.7
Total Phenols	0.001	0.002	-0.015	0.164	0.316	0.7
Flavanoids	0.002	0.000	-0.049	0.214	0.545	1.1
Nonflavanoid_Phenols	0.000	0.002	0.004	-0.025	-0.040	0.1
Proanthocyanins	0.001	-0.007	-0.031	0.082	0.244	0.7
Color Intensity	0.002	0.022	0.097	-0.804	0.536	1.6
Hue	0.000	-0.002	-0.021	0.096	0.064	0.2
OD280/OD315	0.001	-0.002	-0.022	0.220	0.261	0.7
Proline	1.000	0.014	0.004	0.001	-0.004	351.5
Variance	123594.453	194.345	11.424	2.388	1.391	
% Variance	99.830%	0.157%	0.009%	0.002%	0.001%	
Cumulative %	99.830%	99.987%	99.996%	99.998%	99.999%	

The principal components analysis without scaling is trivial for this data set, The first four components are the four variables with the largest variances in the data and account for almost 100% of the total variance in the data.

Principal Components and Orthogonal Least Squares

The weights computed by principal components analysis have an interesting alternate interpretation. Suppose that we wanted to compute fit a linear surface (a straight line for 2-dimensions and a plane for 3-dimensions) to the data points where the objective was to minimize the sum of squared errors measured by the squared orthogonal distances (squared lengths of perpendiculars) from the points to the fitted linear surface. The weights of the first principal component would define the best linear surface that minimizes this sum. The variance of the first principal component expressed as a percentage of the total variation in the data would be the portion of the variability explained by the fit in a manner analogous to R_2 in multiple linear regression. This property can be exploited to find nonlinear structure in high dimensional data by considering perpendicular projections on non-linear surfaces (Hastie and Stuetzle 1989).