9 Principal Components Analysis (PCA)

9.1 Properties of PCA

One of the main uses of PCA in practice is to reduce the dimensionality of the data set. For this we need some guidelines as to how many components should be retained. Obviously, if some, say the last $(p-q)$, eigenvalues of S are zero then it means that the effective dimension of the data is reduced to q.

We have seen that the first principal component has the largest variance. Accordingly, retaining the first q components is equivalent to projecting the data points $x(r)$, $r = 1, \ldots, n$, onto the q-dimensional subspace with maximum variability. We define the **total variance** to be $\sum_{j=1}^{p} s_j^2$, i.e. the sum of the sample variances of all original variables. It is easy to see that

$$\sum_{j=1}^{p} s_j^2 = tr(S) = \sum_{j=1}^{p} \lambda_j = \sum_{j=1}^{p} s_{y_j}^2,$$

We can then say that the jth PC accounts for a proportion $\frac{\lambda_j}{\sum_{j=1}^{p} \lambda_j}$ of the total variation and is typically expressed as percentage. Hence, the first q components account for a proportion $\frac{\sum_{j=1}^{q} \lambda_j}{\sum_{j=1}^{p} \lambda_j}$ of the total variation.

This can be further investigated and used in a variety of ways, the most common being:

1. Aim to retain as many components as necessary so that they account for a given percentage of the total variation. This is sometimes taken to be 80%, but can be different depending on the application. The threshold of 80% is often not realistic if p is very large, and it also makes sense to combine this criterion with the interpretation of the PCs as discussed further, i.e. retain as many components as are meaningful and explain close to 80% of the variance.

   In the wine example, if it is important to explain at least 80% of the variance, we would retain the first five principal components, as we can see from

   ```
    > summary(wine.pca)
   Importance of components:
                            PC1    PC2    PC3     PC4     PC5     PC6      PC7
   ```

```
Standard deviation     2.169 1.5802 1.2025 0.95863 0.92370 0.80103 0.74231

                        PC8     PC9     PC10    PC11    PC12    PC13
                       0.59034 0.53748 0.5009 0.47517 0.41082 0.32152

Proportion of Variance 0.362 0.1921 0.1112 0.07069 0.06563 0.04936 0.04239
                       0.02681 0.02222 0.0193 0.01737 0.01298 0.00795

Cumulative Proportion  0.362 0.5541 0.6653 0.73599 0.80162 0.85098 0.89337
                       0.92018 0.94240 0.9617 0.97907 0.99205 1.00000
```

that the first five principal components explain 80.2% of the variance (while the first four components explain just 73.6%, so are not sufficient).

2. **Scree plot.** The scree test is a graphical method first proposed by Cattell (1966). We can plot λ_i (or % variance) against $i, i = 1, \ldots, p$.

   Cattell's idea is essentially to find a place where the smooth decrease of eigenvalues appears to level off to the right of the plot. To the right of this point, one expects to find only factorial scree. Scree is the geological term referring to the debris which collects on the lower part of a rocky slope.

   In other words, when the data essentially vary in a q-dimensional subspace, i.e. the last $p - q$ components represent '*noise*', then we expect there to be a '*elbow*' pattern in the scree plot. Hence we could choose to retain the first q PCs according to where this elbow appears.

```
> par(mfrow=c(2,1))

> wine.pca <- prcomp(wine[2:14])
> screeplot(wine.pca, type="lines", main="Non-scaled PCA")
> wine.pca <- prcomp(standardisedconcentrations)
> screeplot(wine.pca, type="lines", main="Scaled PCA")
> screeplot(wine.pca, type="lines", main="Scaled PCA")
```

The most obvious change in slope in the scree plot occurs at component 4, which is the '**elbow**' of the scree plot. Hence, we can argue based Fig. that the first three components should be retained.

3. **The Kaiser criterion.** Calculate the mean eigenvalue $\bar{\lambda} = \sum_{i=1}^{p} \lambda_i / p$ and retain all components for which $\lambda_i > \bar{\lambda}$. (Notice that if PCA is applied to the correlation matrix R then $\bar{\lambda} = 1$.) This criterion was proposed by Kaiser (1960), and is probably the one most popular.

   We can check this by finding the variance of each of the principal components:

```
> (wine.pca$sdev)^2
4.7058503 2.4969737 1.4460720 0.9189739 0.8532282
0.6416570 0.5510283 0.3484974 0.2888799 0.2509025
0.2257886 0.1687702 0.1033779
```

The above criteria are in no way to be taken as absolute. They should always be combined with subject matter considerations and with the interpretation of the principal components. Sometimes it is clear that the first few PCs have a useful interpretation whereas the others do not.

Which criterion to use?

- The Kaiser and scree plot criteria have been studied in detail (Browne, 1968; Cattell and Jaspers, 1967; Hakstian, Rogers, and Cattell, 1982; Linn, 1968; Tucker, Koopman and Linn, 1969).

- Note conclusions from the Kaiser method and the scree test differ. It is up to an analyst to decide which route to follow.

- Both Kaiser method and scree plot do quite well under normal conditions, that is, when there are relatively few factors and many cases.

- In practice, an additional important aspect is the extent to which a solution is interpretable. Therefore, one usually examines several solutions with more or fewer factors, and chooses the one that makes the best sense.

9.2 Interpretation of PCA

For the interpretation of a PCA, the main task is to describe the dimensions of most variation in the data. By definition, the single linear transformation with most variation is the first PC, hence we should try to describe the first PC in words:

- is it essentially equal to one of the original variables?
- is it a weighted average of (a subset of) the original variables?
- contrast of (a subset of) the original variables?

Interpreting components individually In considering an individual PC we interpret the corresponding eigenvector $\mathbf{a}_i = (a_{1i}, \ldots, a_{pi})^T$ in terms of the original variables. The greater the absolute value of a_{ji} the more important is the variable $\mathbf{x}_j$ for the ith principal component. Sometimes a cutoff value is used to distinguish the less and more '**important**' variables for a given PC, e.g. those with $|a_{ji}| > 0.7 \times \max_k |a_{ki}|$ are the '**important**' ones. This should not be taken as strict rule.

To obtain the loadings for the first principal component in our analysis of the 13 chemical concentrations in wine samples, we use:

```
> wine.pca$rotation[,1]
      V2           V3           V4           V5           V6
-0.144329395  0.245187580  0.002051061  0.239320405 -0.141992042
      V7           V8           V9           V10          V11
  -0.394660845  -0.422934297  0.298533103 -0.313429488  0.088616705
     V12          V13          V14
-0.296714564 -0.376167411  -0.286752227
```

This means that the first principal component is a linear combination of the variables:

$$-0.144 \times Z_2 + 0.245 \times Z_3 + 0.002 \times Z_4 + 0.239 \times Z_5 - 0.142 \times Z_6 - 0.395 \times Z_7$$
$$-0.423 \times Z_8 + 0.299 \times Z9 - 0.313 \times Z_{10} + 0.089 \times Z_{11} - 0.297 \times Z_{12} - 0.376 \times Z_{13}$$
$$-0.287 \times Z_{14},$$

where $Z_2, Z_3, Z_4, \ldots, Z_{14}$ are the standardised versions of the variables $V_2, V_3, V_4, \ldots, V_{14}$ (i.e. each have mean of 0 and variance of 1).

The first principal component has highest (in absolute value) loadings for variable 8 (0.423), variable 7 (-0.395), variable 13 (-0.376), variable 10 (-0.313), variable 12 (-0.297), variable 14 (-0.287), variable 9 (0.299), variable 3 (0.245), and variable 5 (0.239). The loadings for variables 8, 7, 13, 10, 12 and 14 are negative, while those for variables 9, 3, and 5 are positive. Therefore, an interpretation of the first principal component is that it represents a contrast between the concentrations of variables 8, 7, 13, 10, 12 and 14, and the concentrations of variables 9, 3, and 5.

Similarly, we can obtain the loadings for the second principal:

```
> wine.pca$rotation[,2]
      V2           V3           V4           V5           V6
0.483651548  0.224930935  0.316068814 -0.010590502  0.299634003
      V7           V8           V9           V10          V11
0.065039512   -0.003359812  0.028779488  0.039301722 0.529995672
     V12          V13          V14
-0.279235148 -0.164496193  0.364902832
```

The second principal component has highest loadings for variables 11 (0.530), 2 (0.484), 14 (0.365), 4 (0.316), 6 (0.300), 12 (-0.279), and 3 (0.225). The loadings for variables 11, 2, 14, 4, 6 and 3 are positive, while the loading for variable 12 is negative. Therefore, an interpretation of the second principal component is that it represents a contrast between the concentrations of variables 11, 2, 14, 4, 6 and 3, and the concentration of 12. Since the loadings for variable 11 (0.530) and 2 (0.484) are the largest, the contrast is mainly between the concentrations of variables 11 and 2, and the concentration of variable 12.

Scatterplots of the Principal Components We can draw a scatterplot of the first two principal components, and label the data points with the cultivar that the wine samples come from, by using:

```
> plot(wine.pca$x[,1],wine.pca$x[,2])
> text(wine.pca$x[,1],wine.pca$x[,2], wine$V1, cex=0.7, pos=4, col="red")
```

Scatterplot 9.2 shows that wine samples of cultivar 1 have much lower values of the first principal component than wine samples of cultivar 3. Therefore, the first principal component separates wine samples of cultivars 1 from those of cultivar 3.

We can also see that wine samples of cultivar 2 have much higher absolute values of the second principal component than wine samples of cultivars 1 and 3. Therefore, the second principal component separates samples of cultivar 2 from samples of cultivars 1 and 3.

Hence, the first two principal components are reasonably useful for distinguishing wine samples of the three different cultivars.

Projecting variables onto Principal Components While the above addresses how one individual PC is made up from the original variables, we now ask how a set of (usually adjacent) PCs is jointly made up from the original variables. The general idea is to project the '**old**' variables onto the '**new**' ones. Let $\mathbf{y}_i$ be the i-th PC, i.e. $\mathbf{y}_i = X\mathbf{a}_i$. The projection of '**old**' variable $\mathbf{x}_j$ onto $\mathbf{y}_i$ is given by

$$Pr(\mathbf{x}_j)_{\mathbf{y}_i} = \frac{\mathbf{x}_j^T \mathbf{y}_i}{\mathbf{y}_i^T \mathbf{y}_i}\mathbf{y}_i = a_{ij}\mathbf{y}_i,$$

and the length of this projection is $||Pr(\mathbf{x}_j)_{\mathbf{y}_i}|| = |a_{ji}|\sqrt{(n-1)\lambda_i}$. This is due to the fact that

$$\mathbf{y}_i^T\mathbf{y}_i/(n-1) = \frac{(X\mathbf{a}_i)^T X\mathbf{a}_i}{n-1} = \mathbf{a}_i^T S\mathbf{a}_i = \mathbf{a}_i^T\lambda_i\mathbf{a}_i = \lambda_i\mathbf{a}_i^T\mathbf{a}_i = \lambda_i.$$

(Note that there is a slight abuse of notation here because we talk about eigenvalues of sample covariance matrix S but we do not put $\hat{}$ for λ.) Hence, in a coordinate system of the 1st and 2nd PCs the jth old variable has coordinates $\sqrt{\lambda_i}\mathbf{a}_{j1}$ and $\sqrt{\lambda_i}\mathbf{a}_{j2}$. (Note that we drop constant factor $\sqrt{(n-1)}$ here.) In matrix notation for all variables this is $P = AL^{1/2}$, where A is the matrix of eigenvectors and $L^{1/2}$ is the diagonal matrix with entries $\sqrt{\lambda_i}$.

9.3 Large Sample Inferences for PCA

The eigenvalue-eigenvector pairs of S (or of R) form the basis of principal component analysis.

- The eigenvectors determine the directions of maximum variability.

- The eigenvalues are the variances of the linear combinations.

Because estimates $(\hat{\lambda}, \hat{a})$ are subject to sampling variability, it would be useful to know their sampling distributions so the accuracy of the estimates can be assessed. **(Note that we use $\hat{}$ for $\hat{\lambda}$ and $\hat{a}$ to emphasize that these are finite sample estimates.)**

It might be also of interest to test whether the q smallest eigenvalues are equal in order to decide if the PC's associated with the q smallest eigenvalues only represent random variation with no pattern.

The sampling distributions of the eigenpair $(\hat{\lambda}, \hat{a})$ are difficult to derive so we present results without proof.

Theorem Let $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$. Let $\hat{\lambda} = [\hat{\lambda}_1, \ldots, \hat{\lambda}_p]'$ be an estimate of $\lambda = [\lambda_1, \ldots, \lambda_p]'$, i.e. the p-dimensional vector of unobservable population eigenvalues of Σ. If $(n-1)S$ follows a Wishart distribution and all population eigenvalues λ are distinct, then

1. $$\sqrt{n}(\hat{\lambda} - \lambda) \rightarrow N_p(0, 2\Lambda^2),$$

   where $\Lambda = diag\{\lambda_1, \ldots, \lambda_p\}$.

2. $$\sqrt{n(\hat{\mathbf{a}}_i - \mathbf{a}_i)} \sim N_p(0, E_i),$$

   where
   $$E_i = \lambda_i \sum_{k=1, k\neq i}^{p} \frac{\lambda_k}{(\lambda_k - \lambda_i)^2} \mathbf{a}_k \mathbf{a}_k'.$$

   Note that E_i is not diagonal, and the eigenvectors are not independent.

3. each $\hat{\lambda}$ is distributed independently of the elements of the associated $\hat{\mathbf{a}}_i$.

This result implies that the eigenvalues are independently distributed and that the i-th sample eigenvalue has an approximate distribution that is $N(\lambda_i, 2\lambda_i^2/n)$.

Hence, a large sample $100(1-\alpha)\%$ CI for λ_i is given by

$$\frac{\lambda_i}{1+z_{\alpha/2}\sqrt{2/n}} \leq \lambda_i \leq \frac{\lambda_i}{1-z_{\alpha/2}\sqrt{2/n}},$$

where $z_{\alpha/2}$ is the upper $\alpha/2$-th percentile of $N(0,1)$.

We can use the asymptotic results for $\hat{\lambda}_i$ to test various hypotheses on eigenvalues of Σ, e.g. $H_0 : \lambda_i = \lambda$ for $i = (r+1), \dots, p$.

Bartlett (1947, 1950, 1951) developed a test for the hypothesis that $(p-r)$ smaller eigenvalues of Σ are equal for $0 < r < p-1$. If data support H_0, then there is little interest in using more than r components. Bartlett's statistics is given by

$$M\Big[-\ln(det(S)) + \sum_{i=1}^{r}\ln(\lambda_i) + (p-r)\ln(\lambda)\Big] \sim \chi^2_{df},$$

where

$$\begin{aligned} M &= n - r - \frac{1}{6}\Big(2(p-r)+1+\frac{2}{p-r}\Big) \\ \lambda &= \frac{1}{p-r}\Big(tr(S) + \sum_{i=1}^{r}\lambda_i\Big) \\ df &= \frac{1}{2}(p-r-1)(p-r+2). \end{aligned}$$

(Does it remind you anything?) The statistic was then updated by Anderson (1956) and James (1969). The alternative test is the the Maximum Likelihood Lawley test (Lawley 1940, 1941).

Zwick and Velicer (1986) found Bartlett's test to be highly variable because of its sensitivity to a number of influences (e.g. sample size), and proposed the same limitation for the Maximum Likelihood test. There are many alternative solutions, e.g. the bootstrapped eigenvector-eigenvalue method (Lambert et al. 1990) etc. For more details see discussion by Franklin et al., 1995 and other review articles.

```
> library(nFactors)
```

```
> dim(wine)
[1] 178  14
```

```
> results <- nBartlett(x=(wine.pca$sdev)^2, N= 178, alpha=0.05, details=TRUE)

> results
bartlett anderson   lawley
      12       12       12

> summary(results)
Report For a nFactors Class

Details:

    v    values     bartlett bartlett.chi bartlett.df    bartlett.p anderson.chi anderson.
1   1 4.7058503 0.0004687431   1317.18081          78 2.468617e-224   1364.45112          9
2   2 2.4969737 0.0083793293    818.51684          77 2.116685e-124    851.19376          7
3   3 1.4460720 0.0457953078    525.74931          65  2.794813e-73    548.87611          6
4   4 0.9189739 0.1134222368    369.66566          54  1.330940e-48    387.44153          5
5   5 0.8532282 0.1759942932    293.89388          44  1.122513e-38    309.24006          4
6   6 0.6416570 0.3016933761    201.92098          35  2.321592e-25    213.30525          3
7   7 0.5510283 0.4409771993    137.41557          27  9.453489e-17    145.73966          2
8   8 0.3484974 0.6542874274     70.91353          20  1.291055e-07     75.50912          2
9   9 0.2888799 0.7414789597     49.80156          14  6.591589e-06     53.24131
10 10 0.2509025 0.8046670606     36.04001           9  3.900372e-05     38.68415
11 11 0.2257886 0.8615229736     24.61868           5  1.650510e-04     26.53153
12 12 0.1687702 0.9422645738      9.78268           2  7.511350e-03     10.58551
13 13 0.1033779 1.0000000000      0.00000           0  1.000000e+00      0.00000
      anderson.p  lawley.chi lawley.df      lawley.p
1  1.024102e-226 1317.180809        78 2.468617e-224
2  7.342029e-131  818.595273        66 2.783793e-130
3   1.025375e-77  525.842963        55  4.641806e-78
4   6.184371e-52  369.782101        45  1.311546e-52
```

```
5   1.508354e-41  294.043300        36  3.138041e-42
6   1.915878e-27  202.011086        28  2.852216e-28
7   3.034209e-18  137.489920        21  4.049054e-19
8   2.240254e-08   70.950548        15  3.023194e-09
9   1.730761e-06   49.843235        10  2.852221e-07
10  1.313348e-05   36.073611         6  2.667287e-06
11  7.036230e-05   24.642307         3  1.834130e-05
12  5.027879e-03    9.791089         1  1.753596e-03
13  1.000000e+00    0.000000         0  1.000000e+00


 Number of factors retained by index

bartlett anderson   lawley
      12       12       12
```

Summary PCA is often used in conjunction with other data and statistical procedures, including

- Multiple regression to overcome problems of multicollinearity (use PCs as independent/predictor variables) or to select a sub-set of the original variables.
- Multivariate Analysis of Variance (MANOVA)
- Discriminant analysis: get a lower-dimensional ”look” at structure in data.
- Cluster analysis: Scaling (i.e., PCA) and clustering are often both used when concern is with finding groups of similar objects in a space

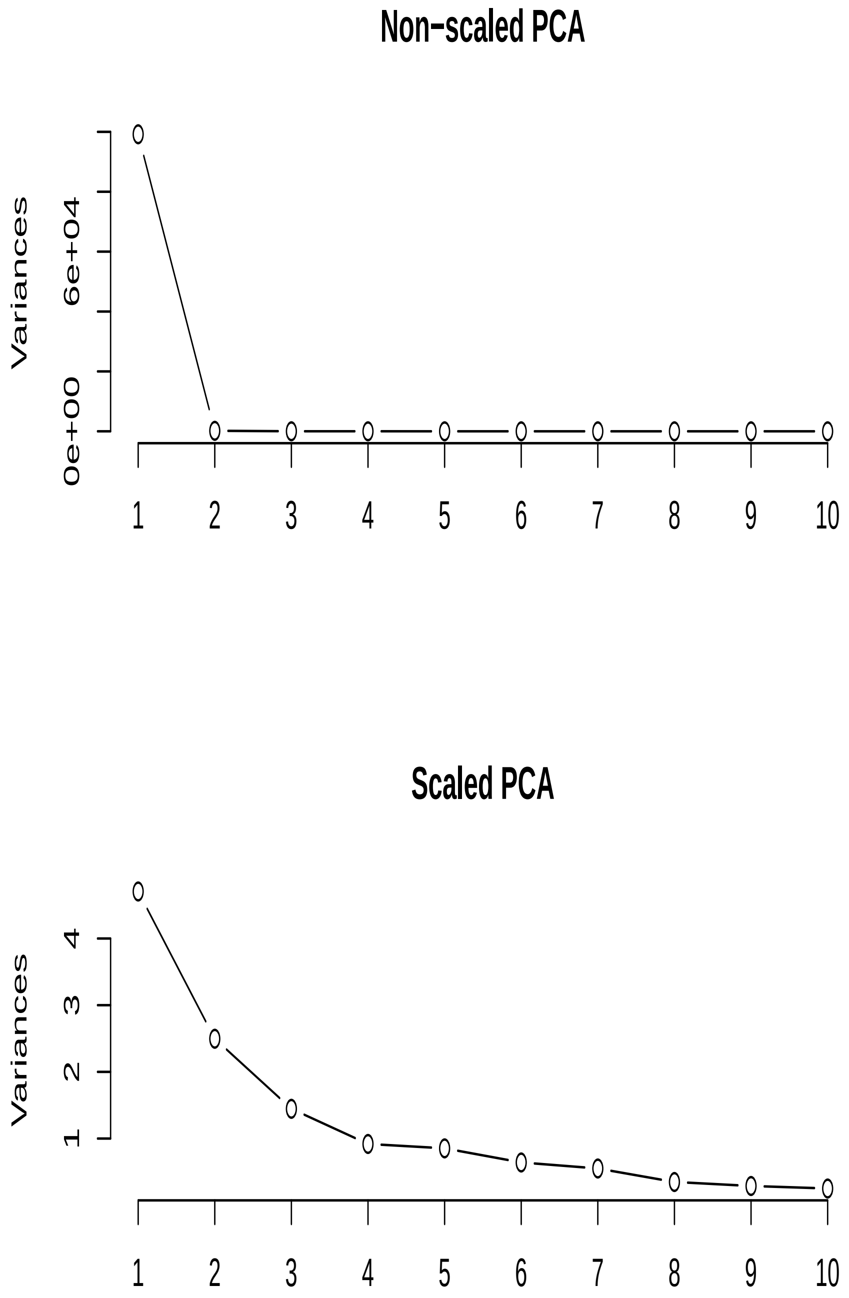

Figure 9.1: Scree plot for PCA for the wine data.

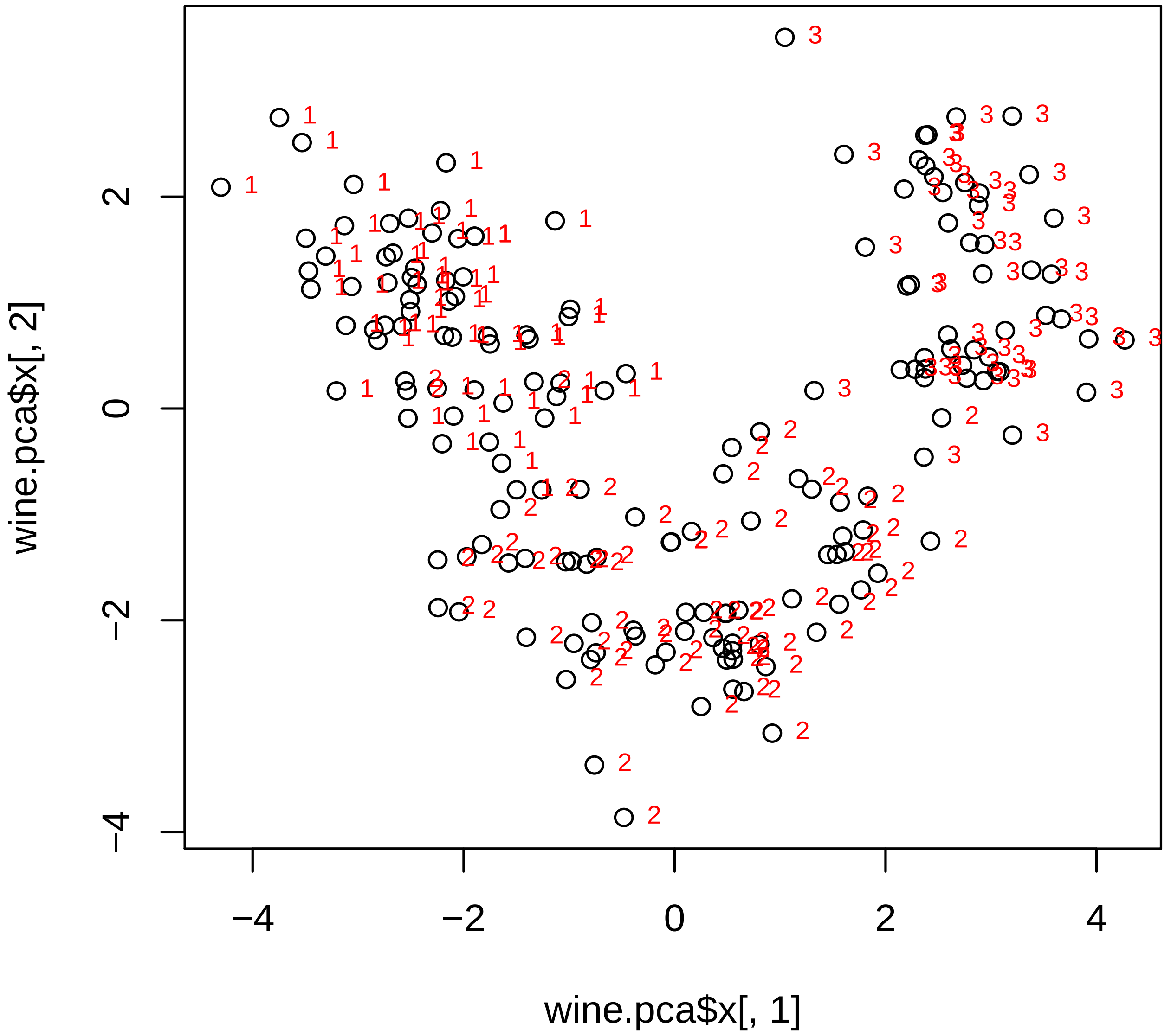

Figure 9.2: Scatter plot for the first two principal components for the wine data.