Principal Components Analysis

Maths and Statistics Help Centre

Usage

- Data reduction tool
- Vital first step in the multivariate analysis of continuous data
- Used to find patterns in high dimensional data
- Expresses data such that similarities and differences are highlighted
- Once the patterns in the data are found, PCA is used to reduce the dimensionality of the data without much loss of information (choosing the right number of principal components is very important)
- If you are using PCA for modelling purposes (either subsequent gradient analyses or regression) then normality is ideal. If it is for data reduction or exploratory purposes, then normality is not a strict requirement

How does PCA work?

- Uses the covariance/correlations of the raw data to define principal components that combine many correlated variables
- The principal components are uncorrelated
- Similar concept to regression – in regression you use one line to describe a relationship between two variables; here the principal component represents the line. Given a point on the line, or a value of the principal component you can discover the values of the variables.

Choosing the right number of principal components

- This is one of the most important parts of PCA. The number you choose needs to be the ones that give you the most information without significant loss of information.
- Scree plots show the eigenvalues. These are used to tell us how important the principal components are.
- When the scree plot plateaus then no more principal components are needed.
- The loadings are a measure of how much each original variable contributes to each of the principal components.

Implementation in R

- princomp(.)
- prcomp(.)

Example

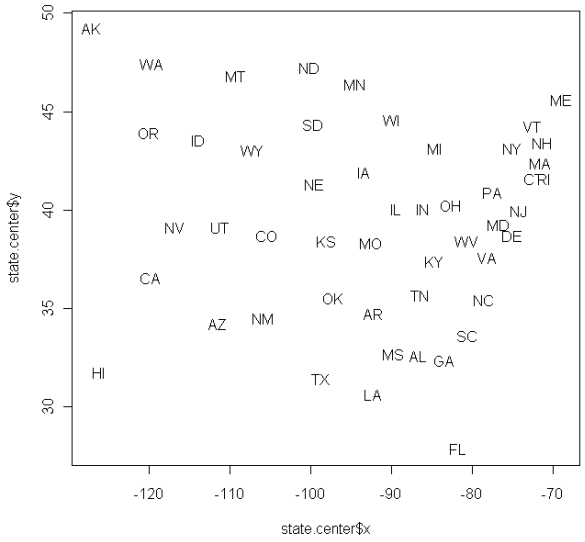
This example uses the built-in R data set *state* with states.x77. The dataset contains 8 indicators about the 50 US states.

```
> data(state)
> ls()
 [1] "state.abb"          "state.area"         "state.center"       "state.division"
 [6] "state.name"         "state.region"       "state.x77"

>state.x77
               Population Income Illiteracy Life Exp Murder HS Grad Frost   Area
Alabama              3615   3624        2.1    69.05   15.1    41.3    20  50708
Alaska                365   6315        1.5    69.31   11.3    66.7   152 566432
Arizona              2212   4530        1.8    70.55    7.8    58.1    15 113417
....
Washington           3559   4864        0.6    71.72    4.3    63.5    32  66570
```

```
West Virginia        1799    3617        1.4     69.48     6.7     41.6    100   24070
Wisconsin            4589    4468        0.7     72.48     3.0     54.5    149   54464
Wyoming               376    4566        0.6     70.29     6.9     62.9    173   97203.
```

```
> plot(state.center,type="n")
> text(state.center,state.abb)
```

Plots where R thinks the states are by centre.



Performing PCA on the states

First perform PCA of the state data using all the information in state.x77

```
state.pca1 <- prcomp(state.x77)
```

Output the PCA summary

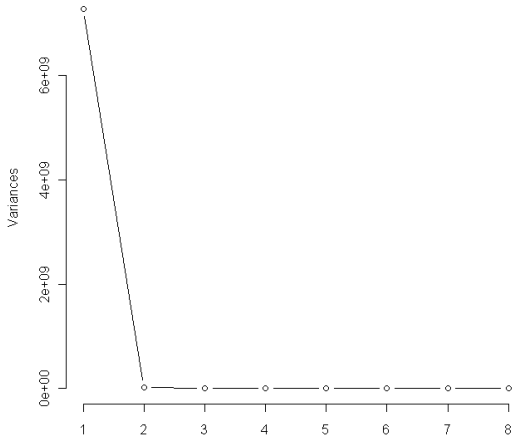
```
>print(state.pca1,digits=3)

Standard deviations:
[1] 8.53e+04 4.47e+03 5.59e+02 4.64e+01 6.04e+00 2.46e+00 6.58e-01 2.90e-01

Rotation:
                PC1       PC2       PC3       PC4       PC5       PC6       PC7       PC8
Population  1.18e-03 -1.00e+00  0.027849 -4.67e-03  3.35e-04  1.39e-04 -5.18e-05 -2.19e-05
Income      2.62e-03 -2.80e-02 -0.999177  2.82e-02 -7.79e-03 -1.12e-04  3.85e-05 -6.29e-05
Illiteracy  5.52e-07 -1.42e-05  0.000584  7.10e-03 -4.05e-02 -3.09e-02  2.55e-02 -9.98e-01
Life Exp   -1.69e-06  1.93e-05 -0.001037 -3.88e-03  1.19e-01  2.86e-01  9.51e-01  1.06e-02
Murder      9.88e-06 -2.79e-04  0.002776  2.82e-02 -2.39e-01 -9.20e-01  3.06e-01  4.62e-02
HS Grad     3.16e-05  1.88e-04 -0.008266 -2.78e-02  9.62e-01 -2.66e-01 -4.08e-02 -3.21e-02
Frost       3.61e-05  3.87e-03 -0.028042 -9.99e-01 -3.45e-02 -1.99e-02  6.25e-03 -4.94e-03
Area        1.00e+00  1.26e-03  0.002583 -3.17e-05 -6.56e-06  1.88e-05 -4.09e-07  1.49e-06
```

We plot the results from our principal component analysis as a scree plot to enable us to decide how many principal components are necessary to best explain the data.

```
>plot(state.pca1,type="l")
```

Principal Components Analysis

Maths and Statistics Help Centre

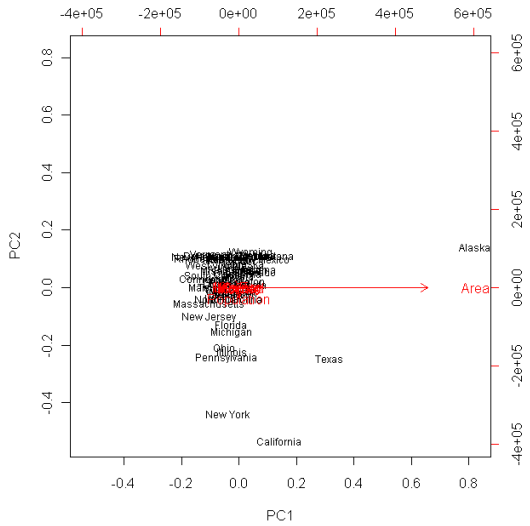



```
state.pca1$sdev[1]^2/(sum((state.pca1$sdev)^2))
[1] 0.9972262
```

The scree plot suggests that the first component explains the majority of the variance (the above calculation shows it to be approximately 99.7%).

Looking at a projection (or biplot) shows us how the components and variables relate, with the magnitude of the arrows representing the magnitude of the effect.

```
biplot(state.pca1,cex=c(0.75,1))
```



The first principal component represents mainly the Area and a bit of the Pop (population) and Income, these are the 3 variables with highest variance.

Looking at the standard deviations of the 8 variables

```
> apply(state.x77,2,sd)

Pop       Income     Illiteracy     Life Exp       Murder      HS Grad         Frost          Area
4.464e+03 6.145e+02  6.095e-01      1.342e+00      3.691e+00   8.077e+00       5.198e+01      8.533e+04
```

This principal component analysis does not seem very informative because the variances are so disparate; we therefore try another method with a scaling factor which we call state.pca2. We do this because when the covariance matrix is unbalanced PCA is very sensitive to the scaling of the original variables; hence we either use the correlation matrix or try scaling. Scaling gives the variables unit variance and is usually advisable.

```
> state.pca2=prcomp(state.x77,scale.=TRUE)
Standard deviations:
[1] 1.8970755 1.2774659 1.0544862 0.8411327 0.6201949 0.5544923 0.3800642 0.3364338
```

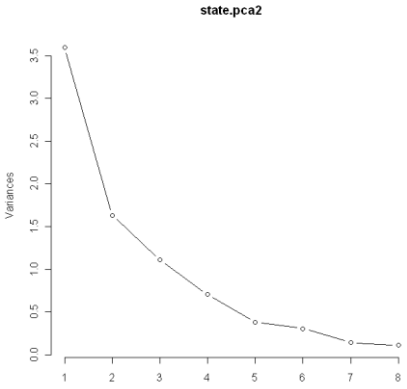
```
Rotation:
                PC1         PC2         PC3         PC4          PC5         PC6          PC7         PC8
Population  0.12642809  0.41087417 -0.65632546 -0.40938555  0.405946365 -0.01065617 -0.062158658 -0.21924645
Income     -0.29882991  0.51897884 -0.10035919 -0.08844658 -0.637586953  0.46177023  0.009104712  0.06029200
Illiteracy  0.46766917  0.05296872  0.07089849  0.35282802  0.003525994  0.38741578 -0.619800310 -0.33868838
Life Exp   -0.41161037 -0.08165611 -0.35993297  0.44256334  0.326599685  0.21908161 -0.256213054  0.52743331
Murder      0.44425672  0.30694934  0.10846751 -0.16560017 -0.128068739 -0.32519611 -0.295043151  0.67825134
HS Grad    -0.42468442  0.29876662  0.04970850  0.23157412 -0.099264551 -0.64464647 -0.393019181 -0.30724183
Frost      -0.35741244 -0.15358409  0.38711447 -0.61865119  0.217363791  0.21268413 -0.472013140  0.02834442
Area       -0.03338461  0.58762446  0.51038499  0.20112550  0.498506338  0.14836054  0.286260213  0.01320320
```

The 1st principal component relates mainly to a combination of illiteracy, life expectancy, murder and HS grad; the 2nd component reflects income, area and population.

```
> summary(state.pca2, digit=3)
Importance of components:
                        PC1   PC2   PC3    PC4    PC5    PC6    PC7    PC8
Standard deviation     1.90 1.277 1.054 0.8411 0.6202 0.5545 0.3801 0.3364
Proportion of Variance 0.45 0.204 0.139 0.0884 0.0481 0.0384 0.0181 0.0141
Cumulative Proportion  0.45 0.654 0.793 0.8813 0.9294 0.9678 0.9859 1.0000
```

A scree plot helps to decide how many components are necessary to explain the data

```
plot(state.pca2,type="l")
```



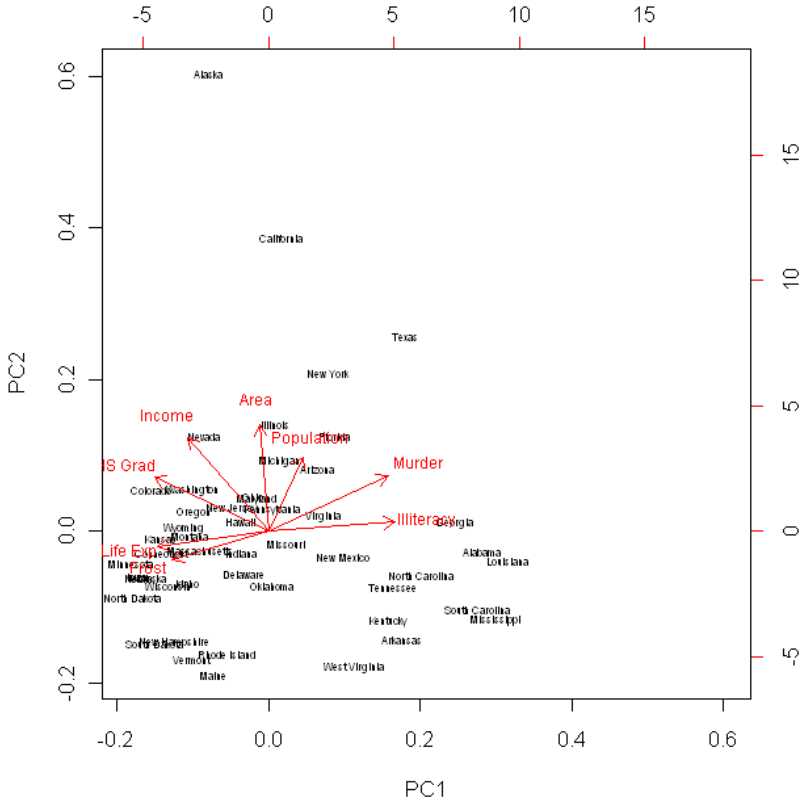
In this case there is no clear plateau point on the plot and so it seems more components are needed, e.g. the first 3 principal components explain close to 80% of the variance. Another way is to decide how many principal components to use is to use only the PCs that have an eigenvalue are greater than 1.

We now plot a biplot which shows a projection of states plus how the variables relate to the components

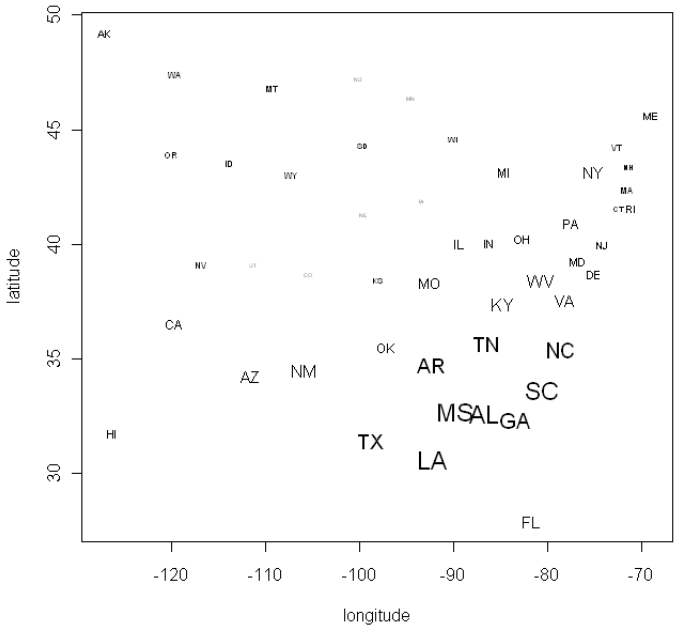
```
> biplot(state.pca2,cex=c(0.5,0.75))
```

Principal Components Analysis

Maths and Statistics Help Centre



PC1 plotted geographically - arguably it is taking into account "Southern-ness". This allows us to understand more about how the principal components are defined.



Plot the second principal component to visualise what it is representing (higher area, population and income).