1 Topic

Bartlett's sphericity test and the KMO index (Kaiser-Mayer-Olkin).

Principal Component Analysis (PCA)[1] is a dimension reduction technique. We obtain a set of factors which summarize, as well as possible, the information available in the data. The factors are linear combinations of the original variables. The approach can handle only quantitative variables.

We have presented the PCA in previous tutorials[2]. In this paper, we describe in details two indicators used for the checking of the interest of the implementation of the PCA on a dataset: the Bartlett's sphericity test and the KMO index. They are directly available in some commercial tools (e.g. SAS or SPSS). Here, we describe the formulas and we show how to program them under R. We compare the obtained results with those of SAS on a dataset.

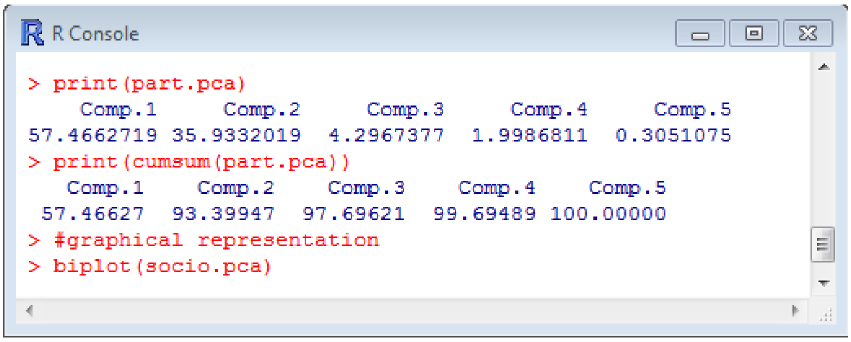
2 Dataset – Principal Component Analysis

Comparing our results on the same dataset with state-of-the-art tools is a good way to validate our program. In this tutorial, we use the formulas available on the SAS and SPSS website. In principle, we should get the same numerical results. We will check it in what follows.

The "socioeconomics.xls" data file contains n = 12 instances and p = 5 variables (POPULATION, SCHOOL, EMPLOYMENT, SERVICES, HOUSEVALUE). We use the following R code to load the dataset and perform the principal component analysis.

```
#importing the data file using the xlsx package
library(xlsx)
socio.data <- read.xlsx(file="socioeconomics.xls", header=T, sheetIndex=1)
#performing the PCA using the princomp command
socio.pca <- princomp(socio.data, cor=T)
#proportion of explained variance of the factors
part.pca <- socio.pca$sdev^2/sum(socio.pca$sdev^2)*100
print(part.pca)
print(cumsum(part.pca))
#graphical representation
biplot(socio.pca)
```

We obtain:

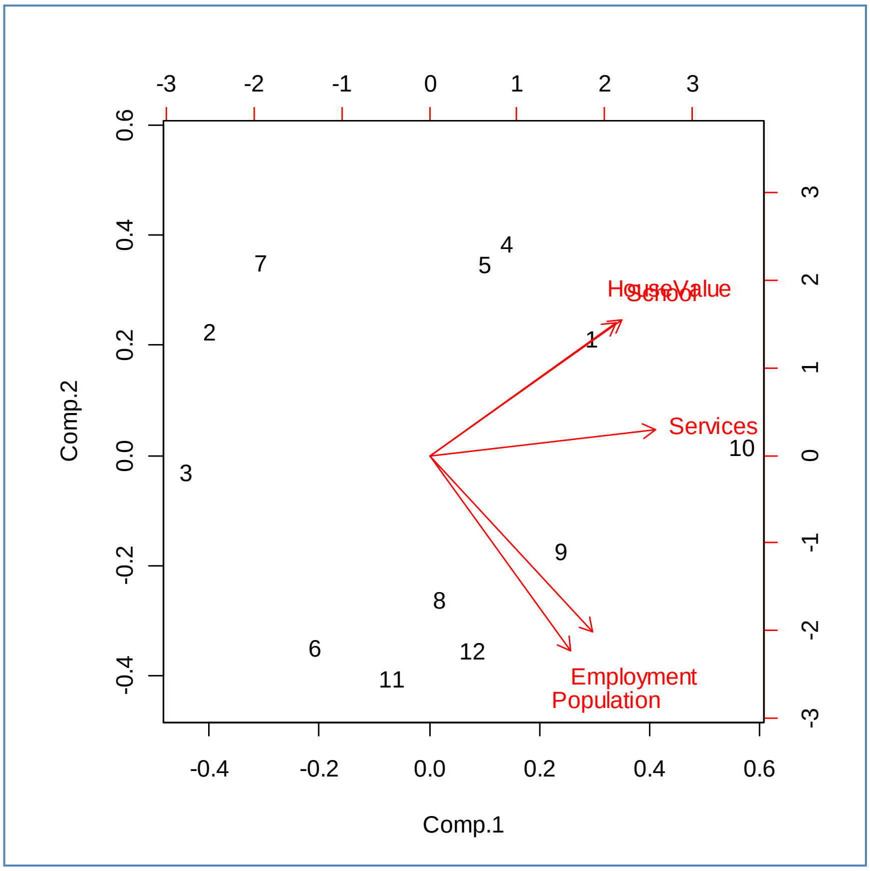


```
> print(part.pca)
    Comp.1     Comp.2     Comp.3     Comp.4     Comp.5 
57.4662719 35.9332019  4.2967377  1.9986811  0.3051075 
> print(cumsum(part.pca))
   Comp.1    Comp.2    Comp.3    Comp.4    Comp.5 
 57.46627  93.39947  97.69621  99.69489 100.00000 
> #graphical representation
> biplot(socio.pca)
```

[1] http://en.wikipedia.org/wiki/Principal_component_analysis

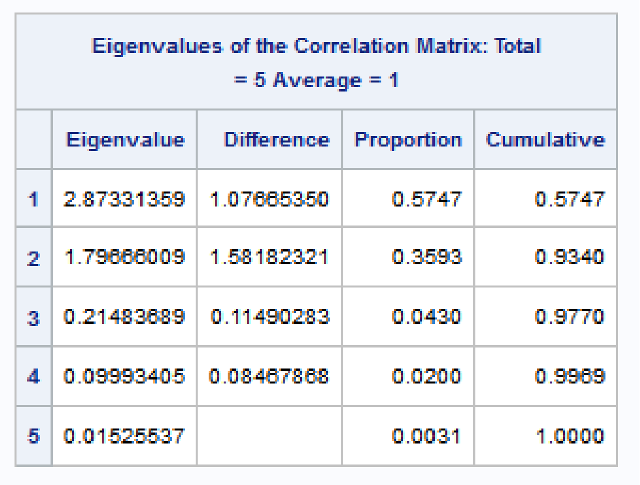
[2] http://data-mining-tutorials.blogspot.fr/2009/04/principal-component-analysis-pca.html

And the following PCA "biplot":



The two first factors represent 93.4% of the available variance. We can consider that we have an accurate picture of the information available in data from these factors.

Using the SAS' FACTOR procedure, we obtain identical results. Here is the eigenvalues table.



Eigenvalues of the Correlation Matrix: Total = 5 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	2.87331359	1.07665350	0.5747	0.5747
2	1.79666009	1.58182321	0.3593	0.9340
3	0.21483689	0.11490283	0.0430	0.9770
4	0.09993405	0.08467868	0.0200	0.9969
5	0.01525537		0.0031	1.0000

3 Bartlett's sphericity test

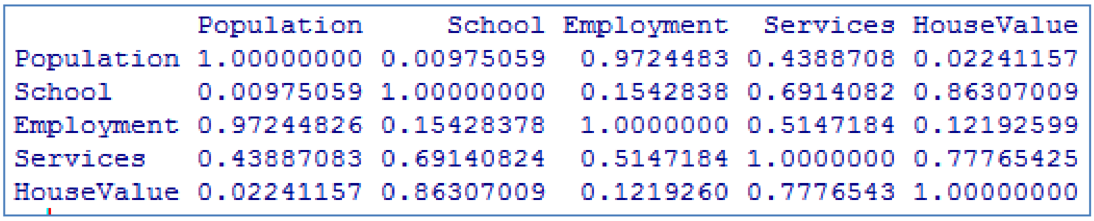
3.1 Calculation of the correlations between the variables

The Bartlett's test compares the observed correlation matrix to the identity matrix. In other words, it checks if there is a certain redundancy between the variables that we can summarize with a few number of factors. If the variables are perfectly correlated, only one factor is sufficient. If they are orthogonal, we need as many factors as variables. In this last case, the correlation matrix is the same as the identity matrix. A simple strategy is to visualize the correlation matrix. If the values outside the main diagonal are often high (in absolute value), some variables are correlated; if most these values are near to zero, the PCA is not really useful.

We calculate the correlation matrix under R:

```
#correlation matrix
R <- cor(socio.data)
print(R)
```

We obtain.



```
           Population      School Employment  Services HouseValue
Population 1.00000000 0.00975059  0.9724483 0.4388708 0.02241157
School     0.00975059 1.00000000  0.1542838 0.6914082 0.86307009
Employment 0.97244826 0.15428378  1.0000000 0.5147184 0.12192599
Services   0.43887083 0.69140824  0.5147184 1.0000000 0.77765425
HouseValue 0.02241157 0.86307009  0.1219260 0.7776543 1.00000000
```

Figure 1 – Correlation matrix

Some variables are correlated (e.g. Population and Employment: 0.97; School and House Value: 0.86). Here, the goal is only to get an overall impression about the redundancy between the variables. We must confirm this with a rigorous statistical procedure.

3.2 Bartlett's sphericity test

The Bartlett's test checks if the observed correlation matrix $\mathbf{R=(r_{ij})_{(p \times p)}}$ diverges significantly from the identity matrix (theoretical matrix under H0: the variables are orthogonal). The PCA can perform a compression of the available information only if we reject the null hypothesis.

In order to measure the overall relation between the variables, we compute the determinant of the correlation matrix |R|. Under H0, |R| = 1; if the variables are highly correlated, we have $|R| \approx 0$.

The Bartlett's test statistic indicates to what extent we deviate from the reference situation |R| = 1. It uses the following formula.

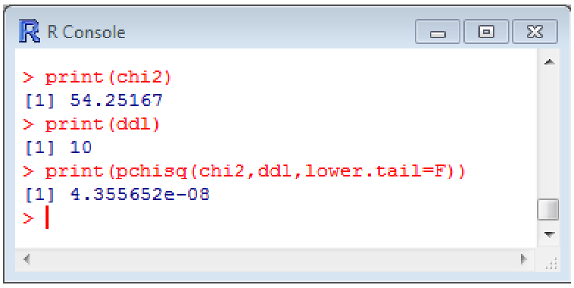
$$\chi^2 = -\left(n - 1 - \frac{2p+5}{6}\right) \times \ln|R|$$

Under H0, it follows a χ^2 distribution with a [p x (p-1) / 2] degree of freedom.

We submit the following program to R:

```
n <- nrow(socio.data)
p <- ncol(socio.data)
chi2 <- -(n-1-(2*p+5)/6)*log(det(R))
ddl <- p*(p-1)/2
print(chi2)
print(ddl)
print(pchisq(chi2,ddl,lower.tail=F))
```

We obtain:

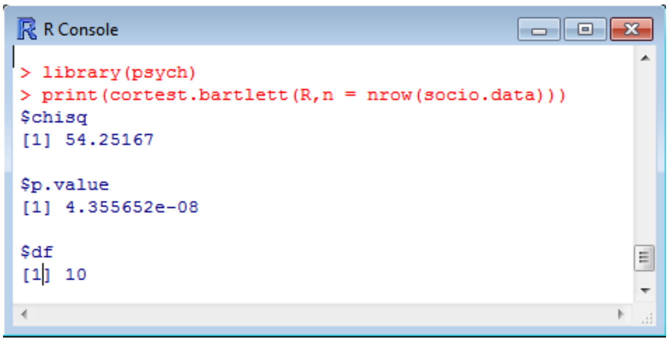


```
> print(chi2)
[1] 54.25167
> print(ddl)
[1] 10
> print(pchisq(chi2,ddl,lower.tail=F))
[1] 4.355652e-08
>
```

We reject the null hypothesis at the 5% level (p-value = 4.35 x 10^{-8} < 0.05). We can perform efficiently a PCA on our dataset.

Note: The Bartlett's test has a strong drawback. It tends to be always statistically significant when the number of instances 'n' increases. Some references advise to use this test only if the ratio 'n:p' (number of instances divided by the number of variables) is lower than 5.

By searching on the web, I found the PSYCH[3] package which proposes the Bartlett's test.



```
> library(psych)
> print(cortest.bartlett(R,n = nrow(socio.data)))
$chisq
[1] 54.25167

$p.value
[1] 4.355652e-08

$df
[1] 10
```

4 KMO Measure of Sampling Adequacy (MSA)

The KMO index has the same goal. It checks if we can factorize efficiently the original variables. But it is based on another idea.

The correlation matrix is always the starting point. We know that the variables are more or less correlated, but the correlation between two variables can be influenced by the others. So, we use the partial correlation in order to measure the relation between two variables by removing the effect of the remaining variables[4]. The KMO index compares the values of correlations between variables and those of the partial correlations. If the KMO index is high (≈ 1), the PCA can act efficiently; if KMO is low (≈ 0), the PCA is not relevant. Some references give a table for the interpretation of the value of the KMO index obtained on a dataset[5].

4.1 Partial correlation matrix

The partial correlation matrix can be obtained from the correlation matrix. We calculate the inverse of this last one $R^{-1} = (v_{ij})$, and we compute the partial correlation $A = (a_{ij})$ as follows:

[3] http://cran.r-project.org/web/packages/psych/index.html

[4] http://en.wikipedia.org/wiki/Partial_correlation

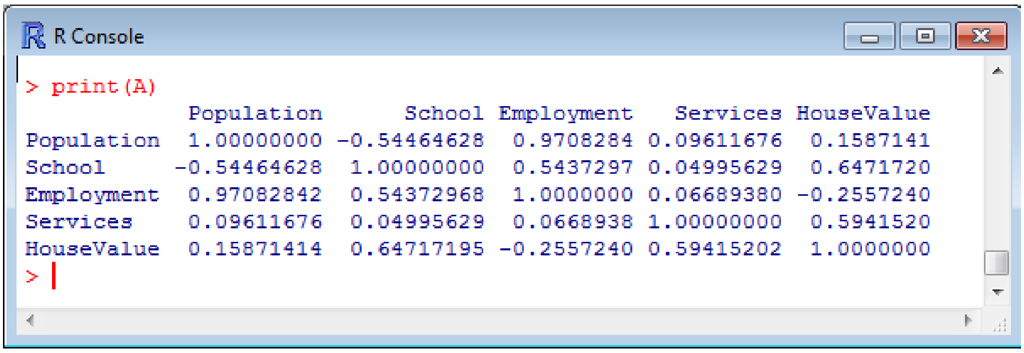
[5] http://peoplelearn.homestead.com/Topic20-FACTORanalysis3a.html

$$a_{ij} = -\frac{v_{ij}}{\sqrt{v_{ii} \times v_{jj}}}$$

Here is a very simplistic program for these calculations under R

```
#inverse of the correlation matrix
invR <- solve(R)
#partial correlation matrix (-1 * spss anti-image matrix, unless the diagonal)
A <- matrix(1,nrow(invR),ncol(invR))
for (i in 1:nrow(invR)){
  for (j in (i+1):ncol(invR)){
    #above the diagonal
    A[i,j] <- -invR[i,j]/sqrt(invR[i,i]*invR[j,j])
    #below the diagonal
    A[j,i] <- A[i,j]
  }
}
colnames(A) <- colnames(socio.data)
rownames(A) <- colnames(socio.data)
print(A)
```

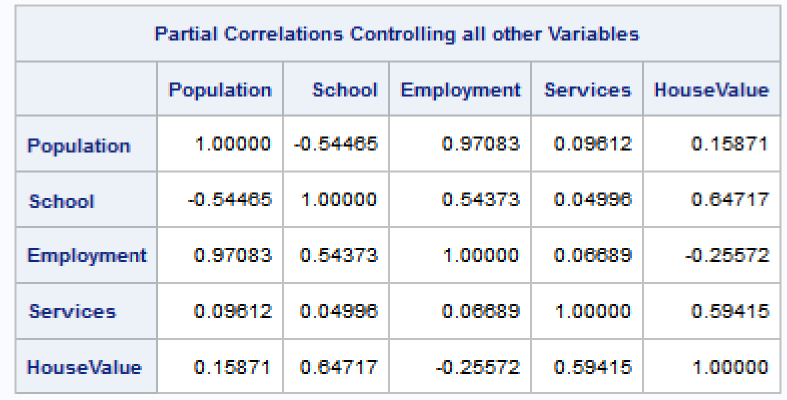
We obtain:



```
> print(A)
            Population      School Employment    Services HouseValue
Population  1.00000000 -0.54464628  0.9708284 0.09611676  0.1587141
School     -0.54464628  1.00000000  0.5437297 0.04995629  0.6471720
Employment  0.97082842  0.54372968  1.0000000 0.06689380 -0.2557240
Services    0.09611676  0.04995629  0.0668938 1.00000000  0.5941520
HouseValue  0.15871414  0.64717195 -0.2557240 0.59415202  1.0000000
> 
```

Figure 2 – Partial correlation matrix

For instance, the correlation between Population and Employment is not influenced by other variables: the partial correlation is very similar to the correlation. SAS provided the same matrix.



Partial Correlations Controlling all other Variables

	Population	School	Employment	Services	HouseValue
Population	1.00000	-0.54465	0.97083	0.09612	0.15871
School	-0.54465	1.00000	0.54373	0.04996	0.64717
Employment	0.97083	0.54373	1.00000	0.06689	-0.25572
Services	0.09612	0.04996	0.06689	1.00000	0.59415
HouseValue	0.15871	0.64717	-0.25572	0.59415	1.00000

SPSS provides the anti-image correlation matrix. This is the partial correlation matrix, but the values on the main diagonal are replaced by the KMO index per variable that we present later.

4.2 Overall KMO index

The overall KMO index is computed as follows.

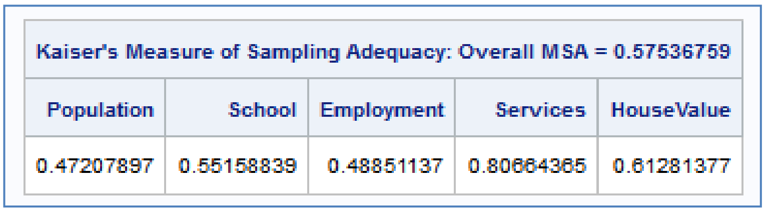
$$KMO = \frac{\sum_i \sum_{j \neq i} r_{ij}^2}{\sum_i \sum_{j \neq i} r_{ij}^2 + \sum_i \sum_{j \neq i} a_{ij}^2}$$

If the partial correlation is near to zero, the PCA can perform efficiently the factorization because the variables are highly related: KMO ≈ 1.

We set the following program under R:

```
kmo.num <- sum(R^2) - sum(diag(R^2))
kmo.denom <- kmo.num + (sum(A^2) - sum(diag(A^2)))
kmo <- kmo.num/kmo.denom
print(kmo)
```

Like SAS, we obtain **KMO = 0.5753676**:



Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.57536759				
Population	School	Employment	Services	HouseValue
0.47207897	0.55158839	0.48851137	0.80664365	0.61281377

Figure 3 – KMO (MSA) index under SAS

With the value 'KMO = 0.575', the degree of common variance in our dataset is rather "mediocre". Into the SAS documentation, one recommends to add variables in the analysis to obtain more reliable results. A commonly used rule is that there should be at least three variables per factor. For our dataset, we have only p = 5 variables for k = 2 factors.

4.3 KMO index per variable

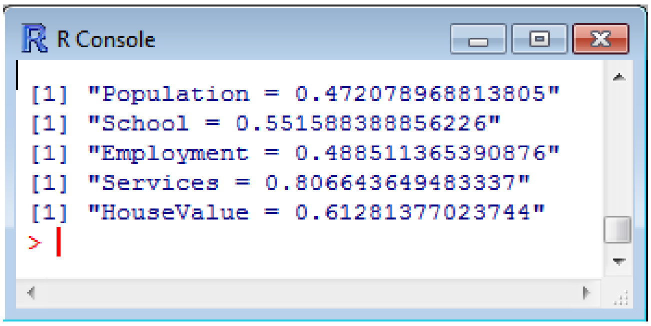
We can compute a KMO index per variable in order to detect those which are not related to the others:

$$KMO_j = \frac{\sum_{i \neq j} r_{ij}^2}{\sum_{i \neq j} r_{ij}^2 + \sum_{i \neq j} a_{ij}^2}$$

We use the following R program:

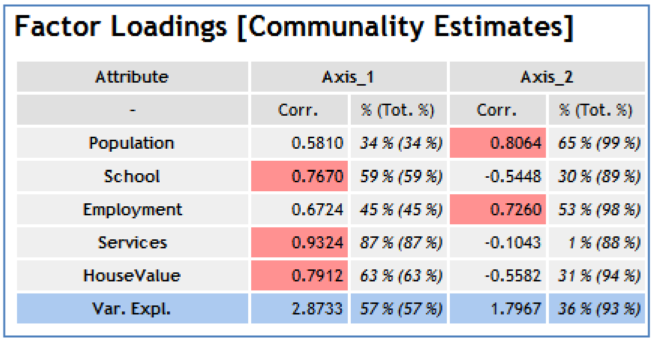
```
#KMO per variable (diagonal of the spss anti-image matrix)
for (j in 1:ncol(socio.data)){
 kmo_j.num <- sum(R[,j]^2) - R[j,j]^2
 kmo_j.denom <- kmo_j.num + (sum(A[,j]^2) - A[j,j]^2)
 kmo_j <- kmo_j.num/kmo_j.denom
 print(paste(colnames(socio.data)[j],"=",kmo_j))
}
```

We obtain (see also Figure 3 for SAS):



```
[1] "Population = 0.472078968813805"
[1] "School = 0.551588388856226"
[1] "Employment = 0.488511365390876"
[1] "Services = 0.806643649483337"
[1] "HouseValue = 0.61281377023744"
>
```

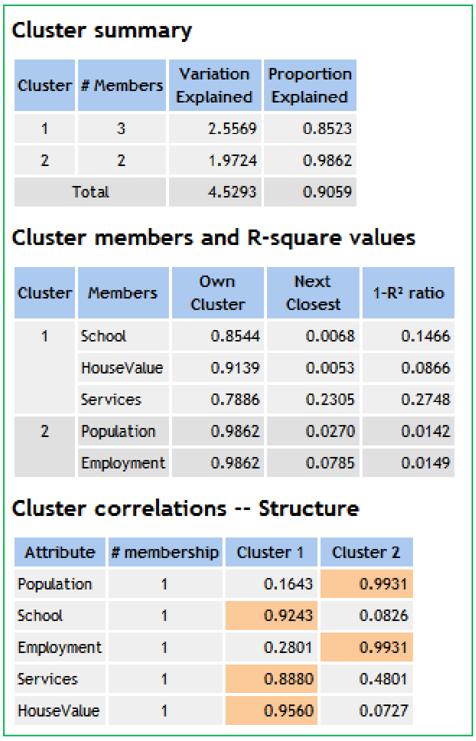
Population (0.472) and Employment (0.488) seems problematic. This is because they are highly correlated (r = 0.94), but not correlated to the other variables (the partial correlation is high also, v = 0.97). It is not really a problem in fact. These variables are not related to the others (which determine the first factor), they define the second factor of the PCA. The output of Tanagra which highlights the high loadings enables to observe this result.



Factor Loadings [Communality Estimates]

Attribute	Axis_1		Axis_2	
-	Corr.	% (Tot. %)	Corr.	% (Tot. %)
Population	0.5810	34 % (34 %)	0.8064	65 % (99 %)
School	0.7670	59 % (59 %)	-0.5448	30 % (89 %)
Employment	0.6724	45 % (45 %)	0.7260	53 % (98 %)
Services	0.9324	87 % (87 %)	-0.1043	1 % (88 %)
HouseValue	0.7912	63 % (63 %)	-0.5582	31 % (94 %)
Var. Expl.	2.8733	57 % (57 %)	1.7967	36 % (93 %)

In a different way, by using a variable clustering process (VARHCA under Tanagra), we observe also these two groups of variables ([School, House Value, Services] vs. [Population, Employment]).



Cluster summary

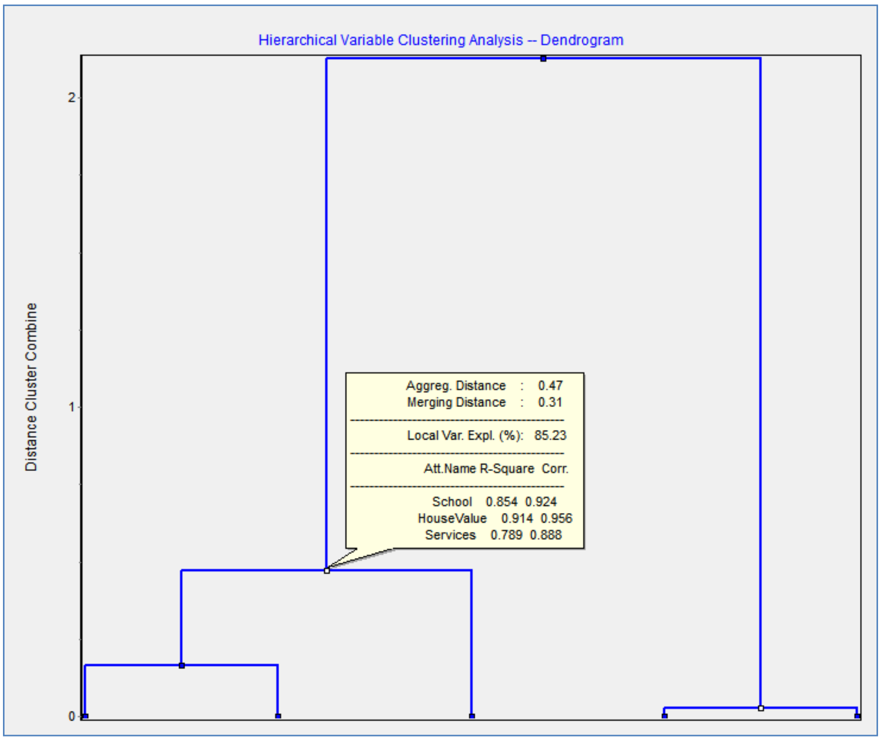
Cluster	# Members	Variation Explained	Proportion Explained
1	3	2.5569	0.8523
2	2	1.9724	0.9862
Total		4.5293	0.9059

Cluster members and R-square values

Cluster	Members	Own Cluster	Next Closest	1-R² ratio
1	School	0.8544	0.0068	0.1466
	HouseValue	0.9139	0.0053	0.0866
	Services	0.7886	0.2305	0.2748
2	Population	0.9862	0.0270	0.0142
	Employment	0.9862	0.0785	0.0149

Cluster correlations -- Structure

Attribute	# membership	Cluster 1	Cluster 2
Population	1	0.1643	0.9931
School	1	0.9243	0.0826
Employment	1	0.2801	0.9931
Services	1	0.8880	0.4801
HouseValue	1	0.9560	0.0727



5 Conclusion

The Bartlett's sphericity test and the KMO index enable to detect if we can or cannot summarize the information provided by the initial variables in a few number of factors. But they do not give indication about the appropriate number of factors. In this tutorial, we show how to compute them with a program written for R. The calculations are feasible only if the correlation matrix is invertible.