

ORIGIN ≡ 1

PRINCIPAL COMPONENTS ANALYSIS

Principal components analysis is perhaps the most popular technique of "ordination" - a very general term connoting a method of data simplification. The object is to plot multivariate data in such a way that hopefully reduces complexity of the problem. In general, n objects (rows in the table below, n=14) are characterized (measured or observed) using p variables (p=5 below). The data are plotted, and the plot "rotated" so as to orient the points along axes of maximum or minimum dispersion. This process is easy to visualize as long as the number of variables p is small. If p=2, for instance, points can be plotted on a piece of paper and directions of maximum and minimum dispersion (called eigenvectors) found by inspection. If p=3, the problem is more complex but visualization is possible in much the same way. In this case points are plotted in 3-D and rotations along three eigenvector axes are possible. When p >3, visualization of the problem becomes abstract, but the underlying principles and mathematics remain the same. After rotation, one inspects the eigenvalues - measures of amount of variation captured by each eigenvector. It is hoped, and often found to be the case, that only a few eigenvectors capture much of the variation in the data. If so, then problem under study might be usefully considered by trying to assign meaning to the perpendicular, and independent, eigenvectors - also called "canonical" directions. By definition, these canonical directions are linear combinations of the original variables associated with coefficients (called "loadings") indicating the angle between vector direction, or equivalently correlation, with the original variables. Data is from RA Johnson & DW Wichern *Applied Multivariate Statistical Analysis 4th Edition 1998*.

Read in Data:

The 1970 Census provided tract information on 5 socioeconomic variables for the Madison, Wisc. area. The data for 14 tracts are listed in J.W. Table 8.5, p. 470.

```
M := READPRN("c:\DATA\Multivariate\T8-5.DAT")
```

```
X := MT
```

Calculating summary statistics:

```
n := cols(X)    n = 14      p := rows(X)    p = 5
```

```
i := 1..n      j := 1..p
```

```
li := 1
```

Mean vector (X_{bar}):

$$X_{\text{bar}} := \frac{1}{n} \cdot X \cdot \mathbf{1} \quad X_{\text{bar}} = \begin{pmatrix} 4.323 \\ 14.014 \\ 1.952 \\ 2.171 \\ 2.454 \end{pmatrix}$$

Mean-Centered Data:

```
Xc,j,i := Xj,i - Xbar,j
```

^ Here the mean of each variable (in X_{bar}) is subtracted from each data value in M (= transpose of X).

Mean-centering is typical of many procedures, including here. Although centering has no effect on the results, all graphs will now include zero - meaning the center of the distribution.

Mean-centering also facilitates comparison with the results of most statistical software that typically reports centered data.

	1	2	3	4	5
1	5.935	14.2	2.265	2.27	2.91
2	1.523	13.1	0.597	0.75	2.62
3	2.599	12.7	1.237	1.11	1.72
4	4.009	15.2	1.649	0.81	3.02
5	4.687	14.7	2.312	2.5	2.22
6	8.044	15.6	3.641	4.51	2.36
7	2.766	13.3	1.244	1.03	1.97
8	6.538	17	2.618	2.39	1.85
9	6.451	12.9	3.147	5.52	2.01
10	3.314	12.2	1.606	2.18	1.82
11	3.777	13	2.119	2.83	1.8
12	1.53	13.8	0.798	0.84	4.25
13	2.768	13.6	1.336	1.75	2.64
14	6.585	14.9	2.763	1.91	3.17

$$X_c^T = \begin{pmatrix} 1.612 & 0.186 & 0.313 & 0.099 & 0.456 \\ -2.8 & -0.914 & -1.355 & -1.421 & 0.166 \\ -1.724 & -1.314 & -0.715 & -1.061 & -0.734 \\ -0.314 & 1.186 & -0.303 & -1.361 & 0.566 \\ 0.364 & 0.686 & 0.36 & 0.329 & -0.234 \\ 3.721 & 1.586 & 1.689 & 2.339 & -0.094 \\ -1.557 & -0.714 & -0.708 & -1.141 & -0.484 \\ 2.215 & 2.986 & 0.666 & 0.219 & -0.604 \\ 2.128 & -1.114 & 1.195 & 3.349 & -0.444 \\ -1.009 & -1.814 & -0.346 & 0.009 & -0.634 \\ -0.546 & -1.014 & 0.167 & 0.659 & -0.654 \\ -2.793 & -0.214 & -1.154 & -1.331 & 1.796 \\ -1.555 & -0.414 & -0.616 & -0.421 & 0.186 \\ 2.262 & 0.886 & 0.811 & -0.261 & 0.716 \end{pmatrix}$$

Variance/covariance matrix (S):

I := identity(n)

$$S := \frac{1}{n-1} \cdot X \cdot \left(I - \frac{1}{n} \cdot \mathbf{1} \cdot \mathbf{1}^T \right) \cdot X^T$$

$$S = \begin{pmatrix} 4.308 & 1.684 & 1.803 & 2.155 & -0.253 \\ 1.684 & 1.767 & 0.588 & 0.178 & 0.176 \\ 1.803 & 0.588 & 0.801 & 1.065 & -0.158 \\ 2.155 & 0.178 & 1.065 & 1.969 & -0.357 \\ -0.253 & 0.176 & -0.158 & -0.357 & 0.504 \end{pmatrix}$$

Sum of Variance in S:

$$\sum_j S_{j,j} = 9.35$$

^ sum of the diagonal of S

$$S_c := \frac{1}{n-1} \cdot X_c \cdot \left(I - \frac{1}{n} \cdot \mathbf{1} \cdot \mathbf{1}^T \right) \cdot X_c^T$$

$$S_c = \begin{pmatrix} 4.308 & 1.684 & 1.803 & 2.155 & -0.253 \\ 1.684 & 1.767 & 0.588 & 0.178 & 0.176 \\ 1.803 & 0.588 & 0.801 & 1.065 & -0.158 \\ 2.155 & 0.178 & 1.065 & 1.969 & -0.357 \\ -0.253 & 0.176 & -0.158 & -0.357 & 0.504 \end{pmatrix}$$

< Note that mean-centering the data has no effect on the variance/covariance matrix. Centering will also have no effect on any calculations involving S below.

^ matrix algebra formula for variance/covariance matrix S or S_c

Eigenvalues & eigenvectors of variance/covariance matrix (S):

Λ := reverse(sort(eigenvals(S)))

E^(j) := eigenvec(S, Λ_j)

$$\Lambda = \begin{pmatrix} 6.931 \\ 1.785 \\ 0.39 \\ 0.23 \\ 0.014 \end{pmatrix}$$

$$E = \begin{pmatrix} 0.78121 & 0.07087 & 0.00366 & -0.54171 & -0.30204 \\ 0.30565 & 0.76387 & -0.16182 & 0.5448 & -0.00928 \\ 0.33445 & -0.08291 & 0.01484 & -0.05102 & 0.93726 \\ 0.42601 & -0.57946 & 0.22045 & 0.63601 & -0.17215 \\ -0.05435 & 0.26236 & 0.96176 & -0.05128 & 0.02458 \end{pmatrix}$$

Testing the Identity SE = ED:

I := identity(p)

D_{j,j} := Λ_j**Diagonal Matrix of Eigenvalues (D):**

$$D = \begin{pmatrix} 6.931 & 0 & 0 & 0 & 0 \\ 0 & 1.785 & 0 & 0 & 0 \\ 0 & 0 & 0.39 & 0 & 0 \\ 0 & 0 & 0 & 0.23 & 0 \\ 0 & 0 & 0 & 0 & 0.014 \end{pmatrix}$$

Sum of Variance in Λ:

$$\sum_j \Lambda_j = 9.35$$

$$S \cdot E = \begin{pmatrix} 5.41461 & 0.12652 & 0.00142 & -0.12434 & -0.00428 \\ 2.11847 & 1.36362 & -0.06305 & 0.12505 & -0.00013 \\ 2.31809 & -0.148 & 0.00578 & -0.01171 & 0.01327 \\ 2.95269 & -1.03442 & 0.0859 & 0.14598 & -0.00244 \\ -0.37673 & 0.46834 & 0.37475 & -0.01177 & 0.00035 \end{pmatrix} \quad E \cdot D = \begin{pmatrix} 5.41461 & 0.12652 & 0.00142 & -0.12434 & -0.00428 \\ 2.11847 & 1.36362 & -0.06305 & 0.12505 & -0.00013 \\ 2.31809 & -0.148 & 0.00578 & -0.01171 & 0.01327 \\ 2.95269 & -1.03442 & 0.0859 & 0.14598 & -0.00244 \\ -0.37673 & 0.46834 & 0.37475 & -0.01177 & 0.00035 \end{pmatrix}$$

Calculating Principal Component Scores:

$Y := E^T \cdot X$ < Scores Y are a linear combinations of X with coefficients E

$Y_c := E^T \cdot X_c$

$$Y_c^T = \begin{pmatrix} 1.438 & 0.293 & 0.441 & -0.749 & -0.201 \\ -3.535 & 0.083 & -0.036 & 0.175 & -0.167 \\ -2.4 & -0.644 & -0.744 & -0.383 & 0.027 \\ -0.595 & 1.846 & 0.046 & -0.063 & 0.048 \\ 0.767 & 0.268 & -0.257 & 0.379 & 0.159 \\ 4.957 & -0.045 & 0.207 & 0.254 & 0.039 \\ -2.132 & -0.063 & -0.618 & -0.211 & -0.002 \\ 2.991 & 2.097 & -0.998 & 0.563 & -0.125 \\ 3.172 & -2.856 & 0.517 & 0.332 & -0.1 \\ -1.421 & -1.6 & -0.323 & -0.386 & -0.02 \\ -0.365 & -1.381 & -0.319 & 0.187 & 0.201 \\ -3.298 & 0.977 & 1.441 & 0.516 & 0.037 \\ -1.737 & -0.083 & 0.138 & 0.371 & -0.027 \\ 2.158 & 1.109 & 0.508 & -0.987 & 0.131 \end{pmatrix}$$

$$Y^T = \begin{pmatrix} 10.54307 & 10.52792 & 1.05666 & 5.70009 & -0.12073 \\ 5.57054 & 10.31795 & 0.57977 & 6.62406 & -0.08673 \\ 6.70519 & 9.59088 & -0.12829 & 6.06572 & 0.10773 \\ 8.51014 & 12.08123 & 0.66259 & 6.38543 & 0.12841 \\ 9.87215 & 10.50321 & 0.35897 & 6.8278 & 0.23908 \\ 14.0629 & 10.19044 & 0.8231 & 6.70301 & 0.11982 \\ 6.97371 & 10.1724 & -0.00186 & 6.23808 & 0.0782 \\ 12.09675 & 12.3326 & -0.382 & 7.01154 & -0.0447 \\ 12.27726 & 7.37897 & 1.13289 & 6.78052 & -0.01945 \\ 7.68473 & 8.63524 & 0.29277 & 6.06258 & 0.06053 \\ 8.74052 & 8.85472 & 0.29668 & 6.63587 & 0.28168 \\ 5.80693 & 11.21198 & 2.05702 & 6.96503 & 0.11763 \\ 7.36805 & 10.15264 & 0.75407 & 6.81931 & 0.05357 \\ 11.26387 & 11.34422 & 1.12385 & 5.46163 & 0.21157 \end{pmatrix}$$

< The "projection" here involves linear combinations of the original data, so mean-centering gives different values that only reflect a shift of each column of data to a center of zero.

Calculating correlation of original variables on the principal components:

$k := 1..p$

$\rho_{j,k} := \frac{E_{j,k} \cdot \sqrt{\Lambda_j}}{\sqrt{S_{j,j}}}$ J.W Result 8.3, p. 429

NOTE:

Be certain that Eigenvalues (Λ) are in rank order and that Eigenvectors (E) are in columns in the same rank order. Otherwise calculations below will take into account the wrong entry in the variance/covariance matrix S .

Eigenvectors
(component values = "loadings")

Correlations with original variables

$$E = \begin{pmatrix} 0.78121 & 0.07087 & 0.00366 & -0.54171 & -0.30204 \\ 0.30565 & 0.76387 & -0.16182 & 0.5448 & -0.00928 \\ 0.33445 & -0.08291 & 0.01484 & -0.05102 & 0.93726 \\ 0.42601 & -0.57946 & 0.22045 & 0.63601 & -0.17215 \\ -0.05435 & 0.26236 & 0.96176 & -0.05128 & 0.02458 \end{pmatrix}$$

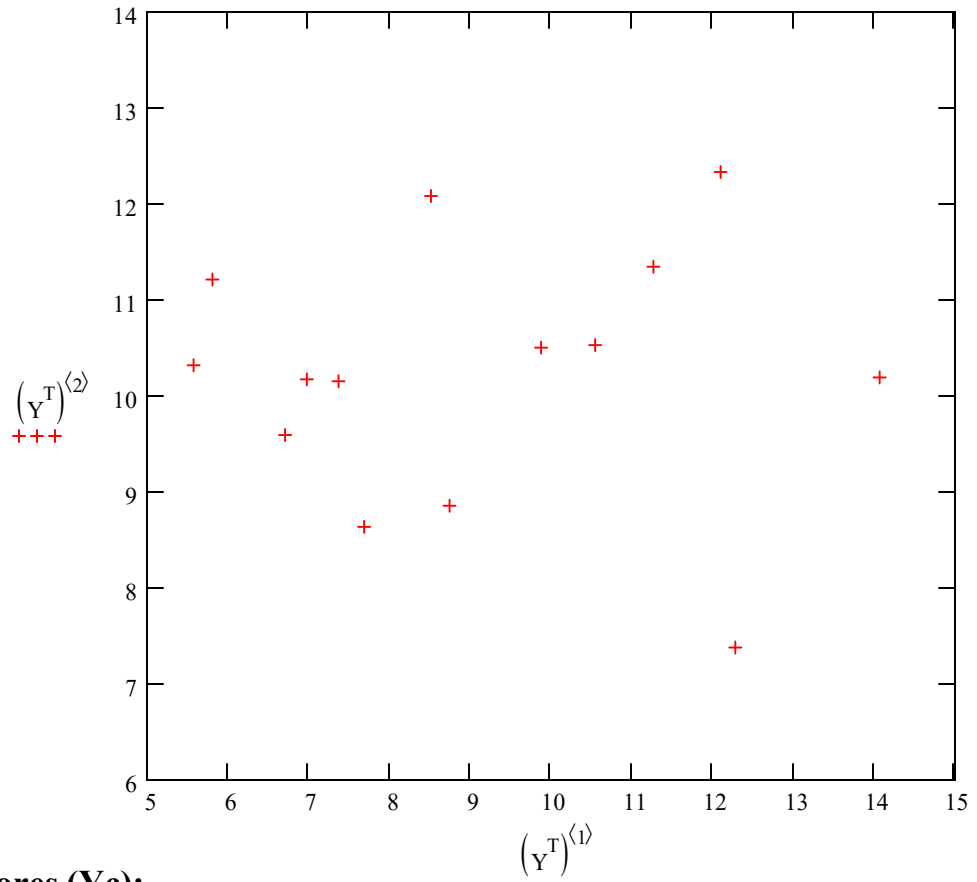
$$\rho^T = \begin{pmatrix} 0.99095 & 0.30717 & 0.23331 & 0.14543 & -0.00911 \\ 0.0899 & 0.76768 & -0.05784 & -0.19782 & 0.04395 \\ 0.00464 & -0.16262 & 0.01035 & 0.07526 & 0.16112 \\ -0.68715 & 0.54752 & -0.03559 & 0.21712 & -0.00859 \\ -0.38313 & -0.00933 & 0.65384 & -0.05877 & 0.00412 \end{pmatrix}$$

^ Original variables - each row
Eigenvectors - each column

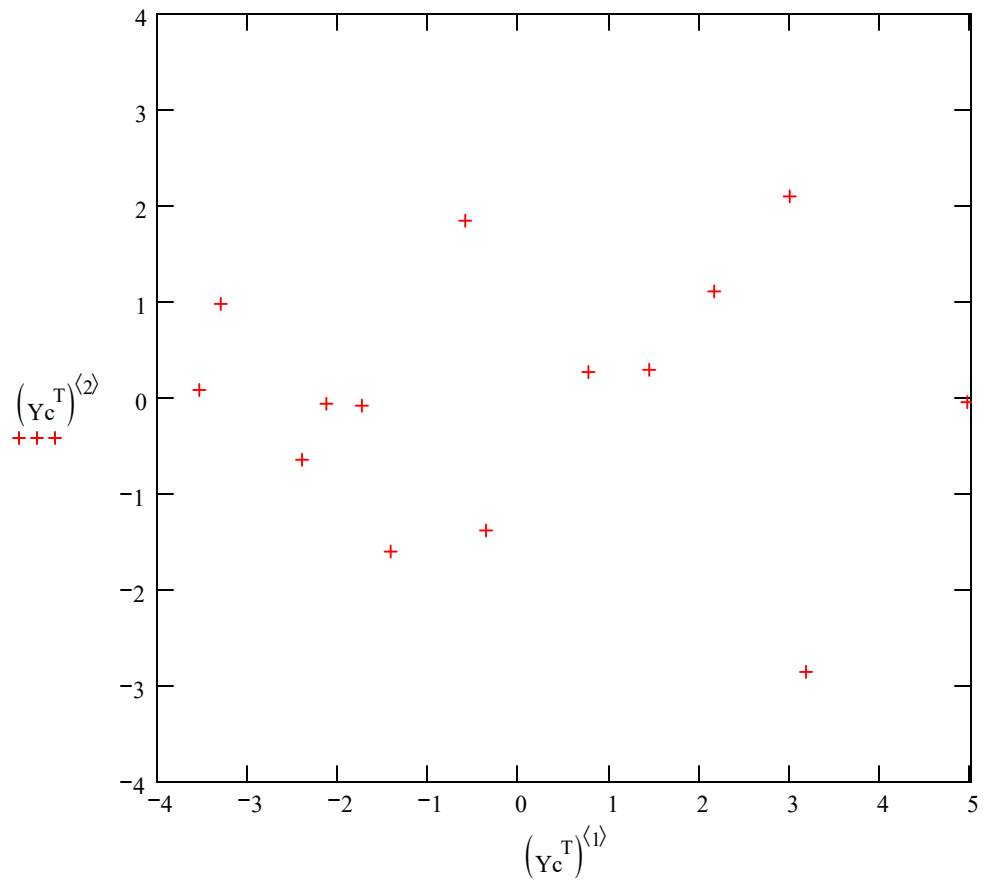
^ Compare with results in chart JW p. 440

^ Note that calculations here only involve eigenvalues & Eigenvectors derived from S, so centering has no effect.

Plot of component scores (Y):



Plot of component scores (Yc):



Fraction Total Variance Explained by each Principal Component:

Eigenvalues and Total Variance:

$$\Lambda = \begin{pmatrix} 6.931 \\ 1.785 \\ 0.39 \\ 0.23 \\ 0.014 \end{pmatrix} \quad \sum_j \Lambda_j = 9.35$$

Standard Deviations:

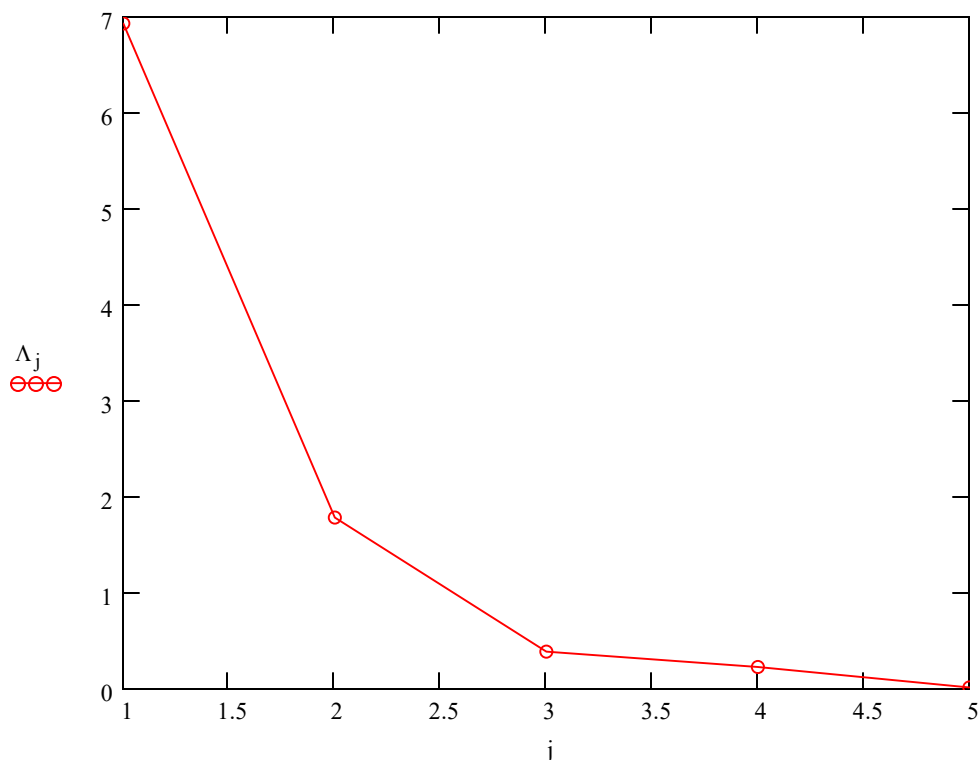
$$\sqrt{\Lambda} = \begin{pmatrix} 2.633 \\ 1.336 \\ 0.624 \\ 0.479 \\ 0.119 \end{pmatrix}$$

Fraction of Variance Explained by each:

$$FE_j := \frac{\Lambda_j}{\sum_j \Lambda_j} \quad FE = \begin{pmatrix} 0.7413 \\ 0.1909 \\ 0.0417 \\ 0.0245 \\ 0.0015 \end{pmatrix}$$

Scree Plot:

$$FE_1 + FE_2 = 0.932$$



Since the first two Principal Components "explain" ~93% of the Total Variance, they would seem to be reasonable descriptions of the data by themselves - thus reducing the data from five dimensions for the original variables to two linear combinations that are the Principal Components.

Note: Use of Principal Components Analysis (PCA) is as close as one can get to "guilt free" statistics!

Choosing the number of Principal Components to use in "data reduction" is a judgement call but doesn't depend on any underlying statistical assumption (such as Normality of the original data). Graphing data in reduced dimensions of the Principal Components is often a first step in interpretation or use of the data with out without further statistical analysis.

A similar analysis might have been conducted using the Correlation Matrix R instead of Variance-Covariance Matrix S. In general, the results will be different. Which to use and interpret is also a judgement call, with no hard-and-fast rules. Greatly different measurement scales among the original variables generally supports use of correlations instead of variance-covariance, but it much depends on whether scale per se is a major part of the problem or something the researcher intends to control. I recommend trying both approaches!

Prototype in R:

#PRINCIPAL COMPONENTS ANALYSIS

```
M=read.table("c:/MultivariateDATA/T8-5.DAT")
```

```
M
```

```
attach(M)
```

```
?prcomp()
```

```
PCS = prcomp(M,retx=TRUE,center=TRUE,scale.=FALSE)
```

```
PCS
```

```
PCS$x
```

```
PCS$center
```

```
PCS$scale.
```

```
> M
```

```

  V1  V2  V3  V4  V5
1  5.935 14.2 2.265 2.27 2.91
2  1.523 13.1 0.597 0.75 2.62
3  2.599 12.7 1.237 1.11 1.72
4  4.009 15.2 1.649 0.81 3.02
5  4.687 14.7 2.312 2.50 2.22
6  8.044 15.6 3.641 4.51 2.36
7  2.766 13.3 1.244 1.03 1.97
8  6.538 17.0 2.618 2.39 1.85
9  6.451 12.9 3.147 5.52 2.01
10 3.314 12.2 1.606 2.18 1.82
11 3.777 13.0 2.119 2.83 1.80
12 1.530 13.8 0.798 0.84 4.25
13 2.768 13.6 1.336 1.75 2.64
14 6.585 14.9 2.763 1.91 3.17
```

```
> PCS
```

```
Standard deviations:
```

```
[1] 2.6326932 1.3360929 0.6242194 0.4790918 0.1189747
```

```
Rotation:
```

```

      PC1  PC2  PC3  PC4  PC5
V1 -0.78120807 0.07087183 -0.003656607 0.54171007 0.302039670
V2 -0.30564856 0.76387277 0.161817438 -0.54479937 0.009279632
V3 -0.33444840 -0.08290788 -0.014841008 0.05101636 -0.937255367
V4 -0.42600795 -0.57945799 -0.220453468 -0.63601254 0.172145212
V5 0.05435431 0.26235528 -0.961759720 0.05127599 -0.024583093
```

```
> PCS$x
```

```

      PC1  PC2  PC3  PC4  PC5
[1,] -1.4376565 0.29260180 -0.44050065 0.74853291 0.201197601
[2,] 3.5348762 0.08263869 0.03638751 -0.17543886 0.167201310
[3,] 2.4002270 -0.64443800 0.74444828 0.38289849 -0.027262233
[4,] 0.5952725 1.84591442 -0.04643003 0.06319255 -0.047938015
[5,] -0.7667385 0.26789299 0.25718390 -0.37918649 -0.158603360
[6,] -4.9574860 -0.04487018 -0.20693717 -0.25439106 -0.039346660
[7,] 2.1317042 -0.06291364 0.61802055 0.21054156 0.002267993
[8,] -2.9913377 2.09728323 0.99815525 -0.56291943 0.125174851
[9,] -3.1718449 -2.85634591 -0.51672992 -0.33189825 0.099924038
[10,] 1.4206830 -1.60007856 0.32338757 0.38604008 0.019946152
[11,] 0.3649005 -1.38059323 0.31947552 -0.18724994 -0.201211729
[12,] 3.2984865 0.97666914 -1.44085808 -0.51641343 -0.037154372
[13,] 1.7373697 -0.08266930 -0.13791241 -0.37069545 0.026902349
[14,] -2.1584560 1.10890855 -0.50769034 0.98698731 -0.131097925
```

```
> PCS$center
```

```

  V1  V2  V3  V4  V5
4.323286 14.014286 1.952286 2.171429 2.454286
```

```
> PCS$scale.
```

```
NULL
```

Note that values of standard deviations, Eigenvectors (Rotation), mean-centered scores (Yc & PCS\$x) and mean values (X_{bar} & PCS\$center) all match.

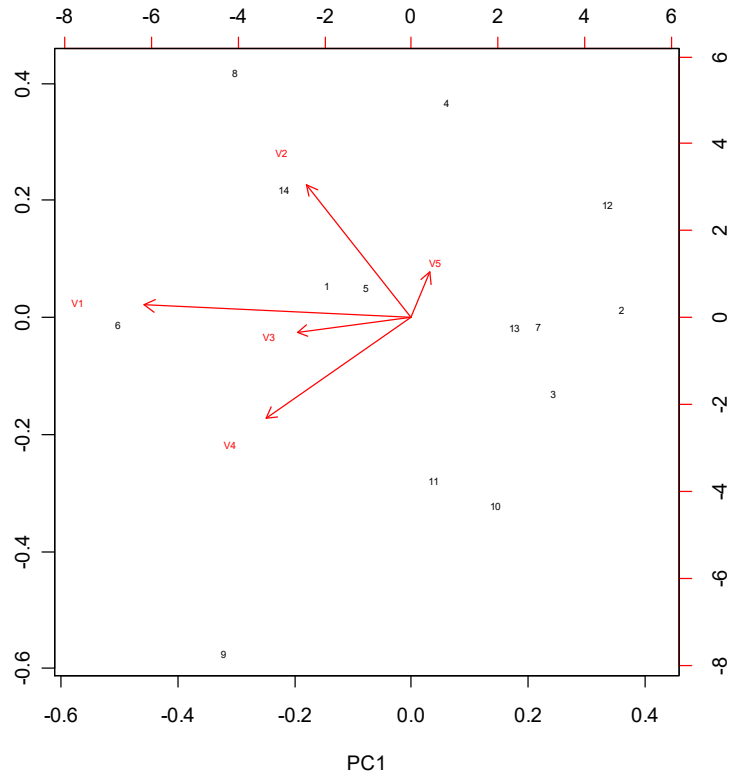
The only differences are seen in the sign of the Eigenvectors and as a result projected scores (+/- for each column). PCA is not defined for +/- so graphs will vary from one implementation of the technique to another. (See also the note in ?prcomp saying the same thing.)

biplot:

```
biplot(PCS,choices=c(1,2),cex=0.5)
biplot(PCS,choices=c(2,3),cex=0.5)
biplot(PCS,choices=c(1,3),cex=0.5)
```

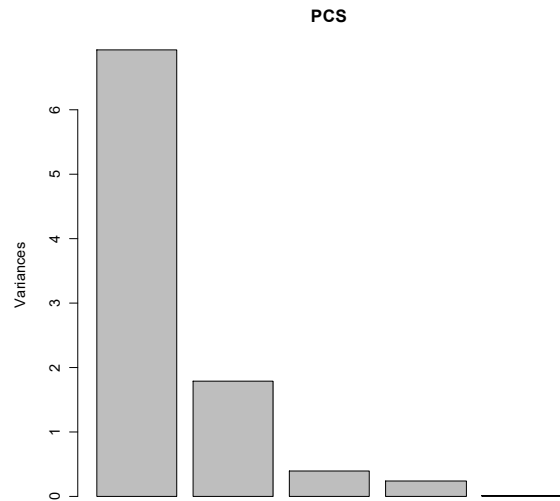
Note: the pattern of scores is the same as above, but directions may be inverted due to +/- uncertainty.

Red vectors (arrows) indicate "directions" of each variable V1-V5 starting at zero and indicated by each ROW in the Rotations list above. Since only PC1 vs PC2 is plotted, only two values for each row are needed to fix the arrowhead.



screplot:

```
plot(PCS)
```



Using the Correlation Matrix:

The correlation matrix **R** is commonly used instead of the variance/covariance matrix **S** in PCA to equalize different scales of measurement typically seen in raw variables (here V1-V5) and for that reason is often the preferred technique. Running PCA involves exactly the same steps. The only difference is that one must first calculate **R** from **S**, or from standardized data.

Calculating Correlation Matrix from the Variance/Covariance Matrix:

$$S = \begin{pmatrix} 4.308 & 1.684 & 1.803 & 2.155 & -0.253 \\ 1.684 & 1.767 & 0.588 & 0.178 & 0.176 \\ 1.803 & 0.588 & 0.801 & 1.065 & -0.158 \\ 2.155 & 0.178 & 1.065 & 1.969 & -0.357 \\ -0.253 & 0.176 & -0.158 & -0.357 & 0.504 \end{pmatrix}$$

< one first constructs a diagonal matrix of standard deviations **D**

$$D_{j,j} := \sqrt{S_{j,j}}$$

$$D = \begin{pmatrix} 2.075 & 0 & 0 & 0 & 0 \\ 0 & 1.329 & 0 & 0 & 0 \\ 0 & 0 & 0.895 & 0 & 0 \\ 0 & 0 & 0 & 1.403 & 0 \\ 0 & 0 & 0 & 0 & 0.71 \end{pmatrix}$$

$$R := D^{-1} \cdot S \cdot D^{-1}$$

^ matrix algebra formula for constructing **R** from **S** & **D** using matrix inverse **D**⁻¹

$$R = \begin{pmatrix} 1 & 0.6102 & 0.9707 & 0.74 & -0.172 \\ 0.6102 & 1 & 0.4943 & 0.0954 & 0.1859 \\ 0.9707 & 0.4943 & 1 & 0.848 & -0.2492 \\ 0.74 & 0.0954 & 0.848 & 1 & -0.358 \\ -0.172 & 0.1859 & -0.2492 & -0.358 & 1 \end{pmatrix}$$

$$\frac{1}{2.075} = 0.482 \quad D^{-1} = \begin{pmatrix} 0.482 & 0 & 0 & 0 & 0 \\ 0 & 0.752 & 0 & 0 & 0 \\ 0 & 0 & 1.118 & 0 & 0 \\ 0 & 0 & 0 & 0.713 & 0 \\ 0 & 0 & 0 & 0 & 1.408 \end{pmatrix}$$

< Note that the matrix inverse of a diagonal matrix **D** is the same as the inverse of each entry in **D**.

Calculating Correlation Matrix from Standardized Data:

$$X_{s,j,i} := \frac{X_{c,j,i}}{D_{j,j}} \quad \text{< Standardized data } X_s \text{ is mean-centered data divided by the associated standard deviation.}$$

$$I := \text{identity}(n)$$

$$R := \frac{1}{n-1} \cdot X_s \cdot \left(I - \frac{1}{n} \cdot \mathbf{1} \cdot \mathbf{1}^T \right) \cdot X_s^T \quad \text{< The correlation matrix is the variance/covariance matrix of standardized data!}$$

$$X_s^T =$$

$$R = \begin{pmatrix} 1 & 0.6102 & 0.9707 & 0.74 & -0.172 \\ 0.6102 & 1 & 0.4943 & 0.0954 & 0.1859 \\ 0.9707 & 0.4943 & 1 & 0.848 & -0.2492 \\ 0.74 & 0.0954 & 0.848 & 1 & -0.358 \\ -0.172 & 0.1859 & -0.2492 & -0.358 & 1 \end{pmatrix}$$

^ same values for **R** as just above.

$$\begin{pmatrix} 0.777 & 0.14 & 0.349 & 0.07 & 0.642 \\ -1.349 & -0.688 & -1.515 & -1.013 & 0.233 \\ -0.831 & -0.989 & -0.799 & -0.756 & -1.034 \\ -0.151 & 0.892 & -0.339 & -0.97 & 0.797 \\ 0.175 & 0.516 & 0.402 & 0.234 & -0.33 \\ 1.793 & 1.193 & 1.887 & 1.666 & -0.133 \\ -0.75 & -0.537 & -0.792 & -0.813 & -0.682 \\ 1.067 & 2.246 & 0.744 & 0.156 & -0.851 \\ 1.025 & -0.838 & 1.335 & 2.386 & -0.626 \\ -0.486 & -1.365 & -0.387 & 0.006 & -0.893 \\ -0.263 & -0.763 & 0.186 & 0.469 & -0.921 \\ -1.346 & -0.161 & -1.29 & -0.949 & 2.528 \\ -0.749 & -0.312 & -0.689 & -0.3 & 0.261 \\ 1.09 & 0.666 & 0.906 & -0.186 & 1.008 \end{pmatrix}$$

Eigenvalues & eigenvectors of correlation matrix (R):

$$\Lambda := \text{reverse}(\text{sort}(\text{eigenvals}(R)))$$

$$E^{(j)} := \text{eigenvec}(R, \Lambda_j)$$

$$\Lambda = \begin{pmatrix} 3.029 \\ 1.291 \\ 0.572 \\ 0.095 \\ 0.012 \end{pmatrix} \quad E = \begin{pmatrix} 0.55836 & 0.13139 & 0.00795 & -0.55055 & -0.60646 \\ 0.31328 & 0.62887 & -0.54903 & 0.45265 & 0.00656 \\ 0.56826 & 0.00426 & 0.11728 & -0.26812 & 0.76904 \\ 0.48662 & -0.30956 & 0.45492 & 0.64798 & -0.20133 \\ -0.17427 & 0.70101 & 0.69122 & -0.01511 & 0.0142 \end{pmatrix}$$

Calculating Principal Component Scores:

$$Y_s := E^T \cdot X_s$$

Correlations of original variables with PCs:

$$k := 1..p$$

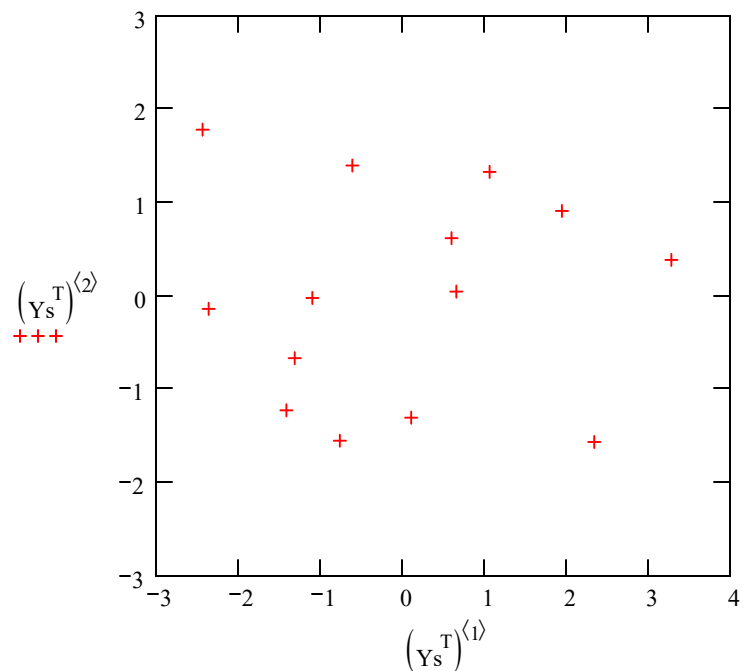
$$\rho_{j,k} := \frac{E_{j,k} \sqrt{\Lambda_j}}{\sqrt{S_{j,j}}}$$

$$\rho^T = \begin{pmatrix} 0.46821 & 0.26776 & 0.4805 & 0.1071 & -0.027 \\ 0.11018 & 0.53749 & 0.0036 & -0.06813 & 0.10863 \\ 0.00666 & -0.46925 & 0.09917 & 0.10012 & 0.10711 \\ -0.46166 & 0.38688 & -0.22671 & 0.14261 & -0.00234 \\ -0.50855 & 0.00561 & 0.65027 & -0.04431 & 0.0022 \end{pmatrix}$$

$$Y_s^T = \begin{pmatrix} 0.5983 & 0.6194 & 0.446 & -0.4222 & -0.2063 \\ -2.363 & -0.1391 & -0.1103 & 0.1778 & -0.1438 \\ -1.4157 & -1.2249 & -0.6163 & -0.2502 & 0.0202 \\ -0.6086 & 1.3982 & -0.4213 & -0.0627 & 0.0437 \\ 0.6593 & 0.0454 & -0.3562 & 0.1859 & 0.1544 \\ 3.2811 & 0.3848 & 0.247 & 0.1287 & 0.0346 \\ -1.314 & -0.6661 & -0.6452 & -0.1346 & -0.0032 \\ 1.9462 & 0.911 & -1.6546 & 0.3434 & -0.1037 \\ 2.3387 & -1.5639 & 1.278 & 0.2538 & -0.0897 \\ -0.7604 & -1.5517 & 0.0854 & -0.2288 & -0.0256 \\ 0.1088 & -1.3047 & 0.0153 & 0.0676 & 0.1903 \\ -2.4373 & 1.7825 & 1.2427 & 0.3609 & 0.05 \\ -1.0991 & -0.0211 & 0.1285 & 0.2576 & -0.0131 \\ 1.0658 & 1.3301 & 0.361 & -0.6772 & 0.0921 \end{pmatrix}$$

^ Correlations with original variables

Plot of component scores (Y):



Fraction Total Variance Explained by each Principal Component:

Eigenvalues and Total Variance:

$$\Lambda = \begin{pmatrix} 3.029 \\ 1.291 \\ 0.572 \\ 0.095 \\ 0.012 \end{pmatrix} \quad \sum_j \Lambda_j = 5$$

Standard Deviations:

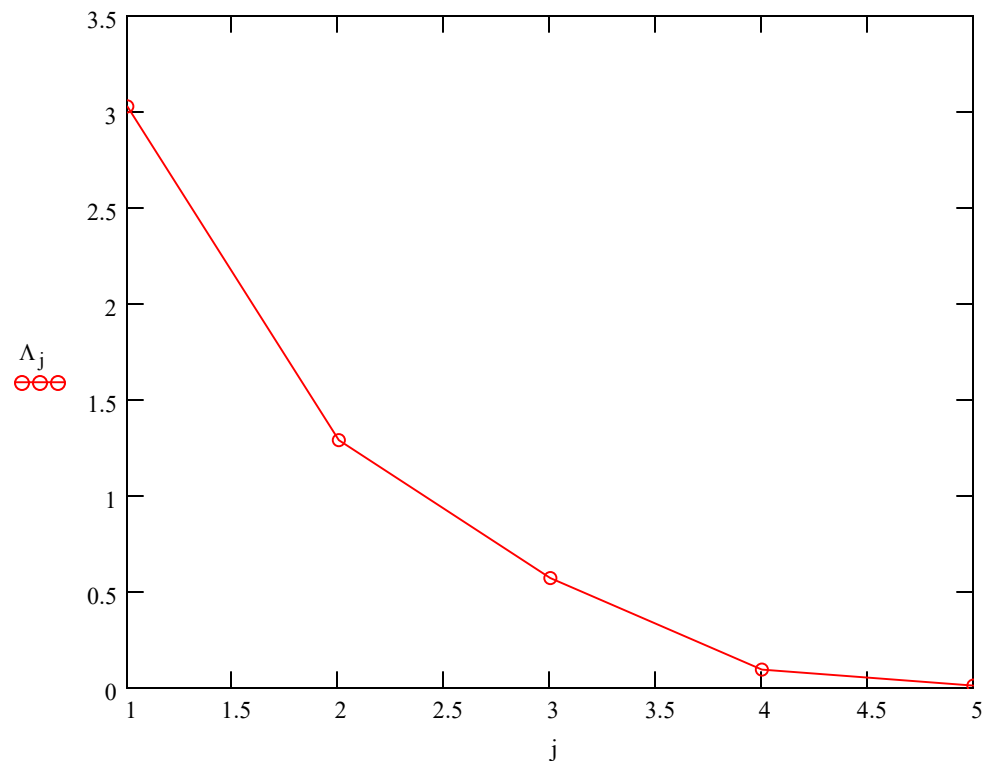
$$\sqrt{\Lambda} = \begin{pmatrix} 1.74 \\ 1.136 \\ 0.757 \\ 0.309 \\ 0.11 \end{pmatrix}$$

Fraction of Variance Explained by each:

$$FE_j := \frac{\Lambda_j}{\sum_j \Lambda_j} \quad FE = \begin{pmatrix} 0.6058 \\ 0.2582 \\ 0.1145 \\ 0.0191 \\ 0.0024 \end{pmatrix}$$

$$FE_1 + FE_2 = 0.864$$

Scree Plot:



Prototype in R:

```
#PCA USING CORRELATION MATRIX
```

```
PCS = prcomp(M, scale = TRUE)
```

```
PCS
```

```
PCS$x
```

```
PCS$center
```

```
PCS$scale.
```

> PCS

Standard deviations:

[1] 1.7403724 1.1362825 0.7566080 0.3088664 0.1100538

Rotation:

	PC1	PC2	PC3	PC4	PC5
V1	-0.5583589	0.131392987	0.007945807	-0.55055321	0.606464575
V2	-0.3132830	0.628872546	-0.549030533	0.45265380	-0.006564747
V3	-0.5682577	0.004262264	0.117280380	-0.26811649	-0.769040874
V4	-0.4866246	-0.309560576	0.454923806	0.64798227	0.201325679
V5	0.1742664	0.701005911	0.691224986	-0.01510711	-0.014203097

> PCS\$x

	PC1	PC2	PC3	PC4	PC5
[1,]	-0.5983117	0.61944480	0.44595617	-0.42218507	0.206299772
[2,]	2.3630458	-0.13910622	-0.11026946	0.17778409	0.143830557
[3,]	1.4157171	-1.22491126	-0.61633603	-0.25023623	-0.020187378
[4,]	0.6086406	1.39823366	-0.42134387	-0.06268978	-0.043651985
[5,]	-0.6592990	0.04537044	-0.35615703	0.18590026	-0.154442903
[6,]	-3.2811300	0.38476758	0.24703862	0.12870817	-0.034614757
[7,]	1.3140407	-0.66607802	-0.64517372	-0.13460249	0.003155881
[8,]	-1.9462339	0.91102625	-1.65457187	0.34338212	0.103700351
[9,]	-2.3387020	-1.56386702	1.27796982	0.25380032	0.089695046
[10,]	0.7603588	-1.55171940	0.08543445	-0.22878447	0.025570341
[11,]	-0.1088036	-1.30466197	0.01530825	0.06761506	-0.190341141
[12,]	2.4373195	1.78246564	1.24265263	0.36091778	-0.050017792
[13,]	1.0991242	-0.02109529	0.12849966	0.25763727	0.013079454
[14,]	-1.0657666	1.33013082	0.36099240	-0.67724703	-0.092075446

biplot(PCS,choices=c(1,2),cex=0.5)

