Correlplot: a collection of functions for plotting correlation matrices

version 1.0-1

Jan Graffelman

Department of Statistics and Operations Research
Universitat Politècnica de Catalunya
Avinguda Diagonal 647, 08028 Barcelona, Spain.
jan.graffelman@upc.edu

October 2013

1 Introduction

This documents gives some instructions on how to create plots of correlation matrices in the statistical environment *R* (R Core Team, 2013) using the package **Correlplot**. The outline of this guide is as follows: Section 2 explains how to install this package and Section 3 shows how to use the functions of the package for creating pictures like correlograms and biplots for a given correlation matrix. Computation of goodness-of-fit statistics is also illustrated. If you appreciate this software then please cite the following paper in your work:

Graffelman, J. (2013) Linear-angle correlation plots: new graphs for revealing correlation structure. *Journal of Computational and Graphical Statistics*, **22**(1) pp. 92-106. (click here to access the paper).

2 Installation

The package **Correlplot** can be installed in *R* by typing:

```
install.packages("Correlplot")
library("Correlplot")
```

This instruction will make, among others, the functions `correlogram`, `linangplot`, `pco` and `pfa` available. The correlation matrices used in the paper cited above are included in the package, and can be accessed with the `data` instruction. The instruction `data(package="Correlplot")` will give a list of all correlation and data matrices available in the package.

3 Plots of a correlation matrix

In this section we indicate how to create different plots of a correlation matrix, and how to obtain the goodness-of-fit of the displays. We will subsequently treat:

- The linear-angle correlation scatterplot (subsection 3.1).

- The principal component analysis (PCA) biplot of a correlation matrix (subsection 3.2).
- The principal factor analysis biplot (PFA) of a correlation matrix (subsection 3.3).
- Correlograms of the correlation matrix (subsection 3.4).

3.1 Linear-angle correlation scatterplot

An ordinary scatterplot represents two variables on orthogonal axes. A principal component analysis of two variables gives a biplot with two axes representing the two variables. If the correlation between the two variables is non-zero, then these two variable axes will be oblique. In the latter plot the angle (α) sustending the two biplot axes satisfies $\cos(\alpha) = r_{xy}$, where r_{xy} is the sample correlation coefficient between the two variables. We first illustrate this plot with Pearson and Lee's classical dataset on heights of daughters and mothers, taken from Weisberg (2005).

```
> data(PearsonLee)
> Mheight <- PearsonLee$Mheight
> Dheight <- PearsonLee$Dheight
> plot(Mheight,Dheight,asp=1,xlab="Mother's height (cm)",
+      ylab="Daughter's height (cm)",pch=19,cex=0.05)
```

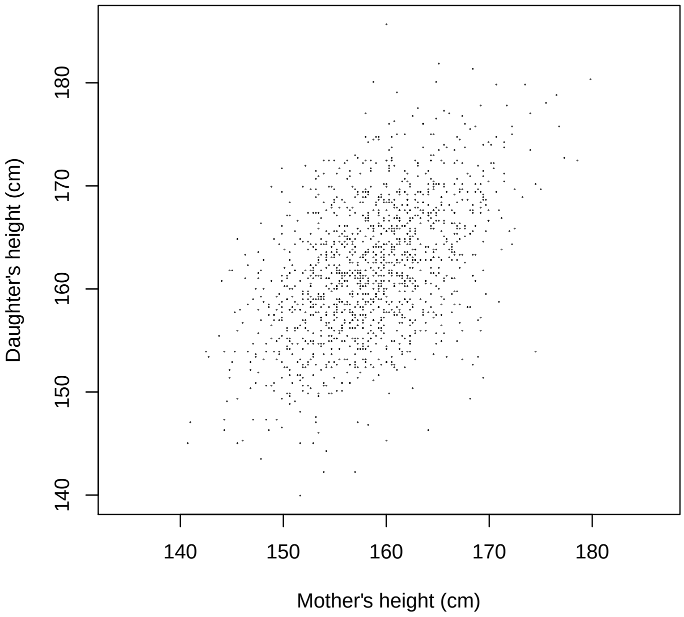

Figure 1: Scatter plot of daughter's height versus mother's height for 1375 pairs.

A scatterplot of the data is shown in Figure 1. Next, we create a PCA biplot of the correlation matrix, shown in Figure 2.

```
> X <- cbind(Mheight,Dheight)
> n <- nrow(X)
> Xt <- scale(X)/sqrt(n-1)
> res.svd <- svd(Xt)
> Fs <- sqrt(n-1)*res.svd$u
> Gp <- res.svd$v%*%diag(res.svd$d)
> plot(Fs[,1],Fs[,2],asp=1,pch=19,cex=0.05,xlab="First principal component",
+      ylab="Second principal component")
> arrows(0,0,3*Gp[,1],3*Gp[,2],col="red",lwd=2)
> textxy(3*Gp[,1],3*Gp[,2],colnames(X),cex=1)
```

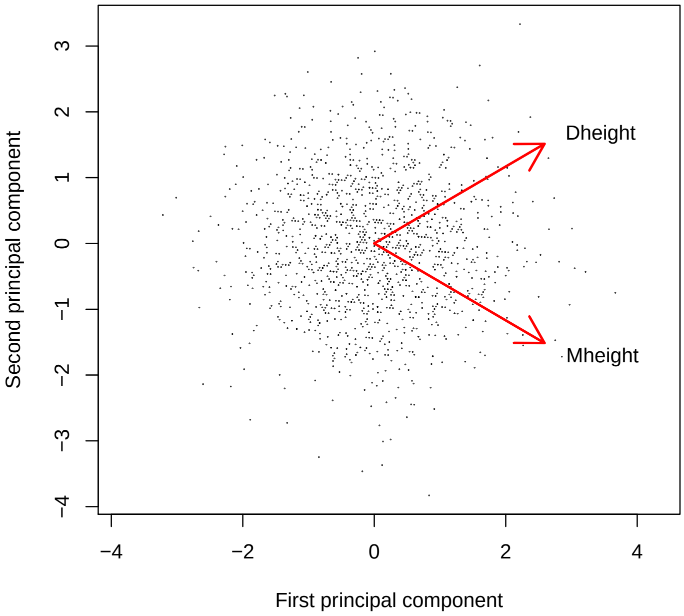

Figure 2: PCA biplot of the mother-daughter height data.

The correlation between the two variable is 0.491. The angle between the two vectors in 60.61°, and we verify $cos(60.61) = 0.491$.

Finally, we call `linangplot` in order to create the linear angle scatterplot shown in Figure 3

```
> lin.out <- linangplot(Mheight,Dheight,cex=0.05)
```

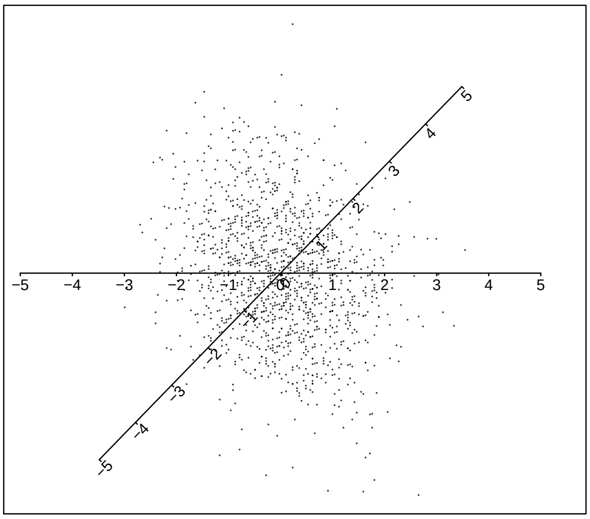

Figure 3: Linear angle correlation scatterplot of the mother-daughter height data.

In the latter plot, the angle between the two axis is 45.84° degrees. In this plot the relation between the correlation and the angle is linear ($r_{xy} = 1 - \frac{2}{\pi}\alpha$), and we have $1 - \frac{2}{\pi}0.8 = 0.491$. Because the rank of the correlation matrix in these examples is 2, Figures 2 and 3 both represent the correlation coefficient without error.

3.2 PCA biplot of the correlation matrix

We use a larger, classical data set of 5 variables with grades of students on 5 exams (Mardia et al., 1979) to illustrate PCA biplots, PFA biplots and correlograms, as well as the corresponding goodness-of fit calculations. The sample correlation matrix is given in Table 1.

The PCA biplot of the correlation matrix can be obtained from a PCA of the data matrix, or, as shown here, directly from the spectral decomposition of the correlation matrix.

	Mec(c)	Vec(c)	Alg(o)	Ana(o)	Sta(o)
Mec(c)	1.000	0.553	0.547	0.409	0.389
Vec(c)	0.553	1.000	0.610	0.485	0.436
Alg(o)	0.547	0.610	1.000	0.711	0.665
Ana(o)	0.409	0.485	0.711	1.000	0.607
Sta(o)	0.389	0.436	0.665	0.607	1.000

Table 1: Correlation matrix for student grades on 5 subjects (Mec=Mecanics,Vec=Vectors,Alg=Algebra,Ana=Analysis,Sta=Statistics).

```
> data(students)
> R <- cor(students)
> out.eigen <- eigen(R)
> V <- out.eigen$vectors
> D <- diag(out.eigen$values)
> F <- V%*%sqrt(D)
> plot(F[,1],F[,2],pch=19,asp=1,xlim=c(-1,1),ylim=c(-1,1))
> origin()
> arrows(0,0,F[,1],F[,2])
> textxy(F[,1],F[,2],colnames(R),cex=1)
```

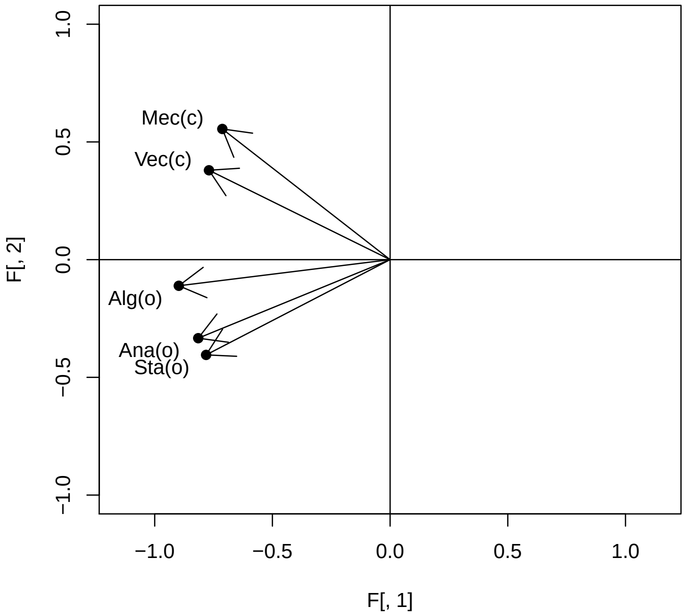

Figure 4: PCA Biplot of a correlation matrix of student grades on 5 exams.

The plot shows sharp angles for all exams, revealing positive correlations. The two dimensional approximation ($\hat{\mathbf{R}}$) made by PCA to the correlation matrix by using cosines is shown in Table 2.

	Mec(c)	Vec(c)	Alg(o)	Ana(o)	Sta(o)
Mec(c)	1.000	0.979	0.708	0.497	0.418
Vec(c)	0.979	1.000	0.836	0.662	0.593
Alg(o)	0.708	0.836	1.000	0.965	0.938
Ana(o)	0.497	0.662	0.965	1.000	0.996
Sta(o)	0.418	0.593	0.938	0.996	1.000

Table 2: Least squares approximation by cosines to the correlation matrix obtained by PCA.

The amount of error in the representation of the correlation matrix is expressed as the root mean squared error (RMSE) of the below-diagonal elements of $\mathbf{R}$, and is for this PCA plot 0.093 if correlations are approximated by scalar products, and 0.248 if correlations are approximated by cosines.

3.3 PFA biplots of a correlation matrix

Principal factor analysis can be performed by the function `pfa` of package **Correlplot**. This function is an adaptation of matlab code provided by Albert Satorra (1998).

```
> out.pfa <- pfa(students)
```

```
Initial communalities
[1] 0.3764142 0.4451224 0.6713576 0.5408636 0.4793185
Final communalities
[1] 0.5194946 0.5924286 0.8115381 0.6480277 0.5692397
9  iterations till convergence
Specific variances:
[1] 0.4805054 0.4075714 0.1884619 0.3519723 0.4307603
Variance explained by each factor
[1] 2.8282754 0.3124535
Loadings:
           [,1]        [,2]
[1,] -0.6400750 -0.33135873
[2,] -0.7103202 -0.29643508
[3,] -0.8966722  0.08670088
[4,] -0.7705227  0.23307202
[5,] -0.7185390  0.23009016
```

```
> L <- out.pfa$La
```

The corresponding plot of the correlation matrix is shown in Figure 5. The approximation to the correlation matrix is given by

```
> Rhatpfa <- L[,1:2]%*%t(L[,1:2])
```

and is shown in Table 3 below.

```
> plot(L[,1],L[,2],pch=19,asp=1,xlim=c(-1,1),ylim=c(-1,1))
> origin()
> arrows(0,0,L[,1],L[,2])
> text(L[,1],L[,2],colnames(students),cex=1,pos=c(1,2,2,2,3))
```

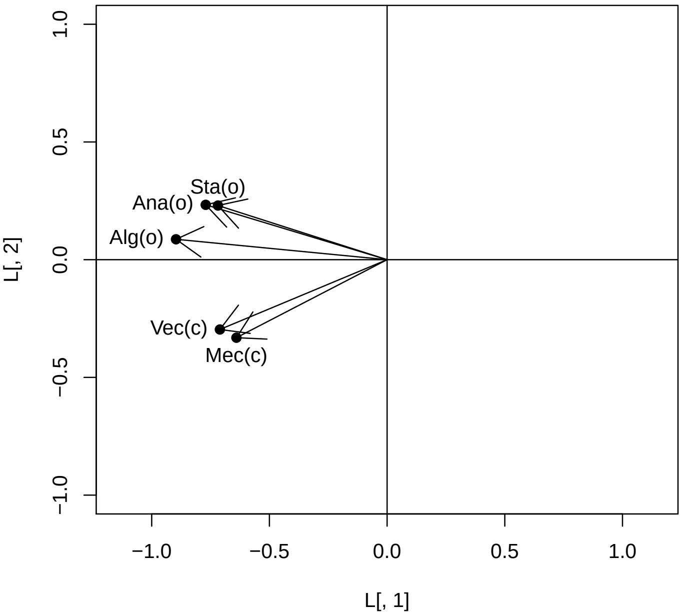

Figure 5: Plot of the correlation matrix of student grades on 5 exams obtained by principal factor analysis.

	Mec(c)	Vec(c)	Alg(o)	Ana(o)	Sta(o)
Mec(c)	0.519	0.553	0.545	0.416	0.384
Vec(c)	0.553	0.592	0.611	0.478	0.442
Alg(o)	0.545	0.611	0.812	0.711	0.664
Ana(o)	0.416	0.478	0.711	0.648	0.607
Sta(o)	0.384	0.442	0.664	0.607	0.569

Table 3: Least squares approximation, using scalar products, to the correlation matrix obtained by PFA.

The RMSE of the plot obtained by PFA is 0.004, and this is lower than the RMSE obtained by PCA.

3.4 Correlograms of a correlation matrix

The correlogram proposed by Trosset (2005) can be obtained by the instructions:

```
> correlogram(R,labs=colnames(R),main="",
+             xlim=c(-1.3,1.3),ylim=c(-1.3,1.3))
```

and is shown in Figure 6

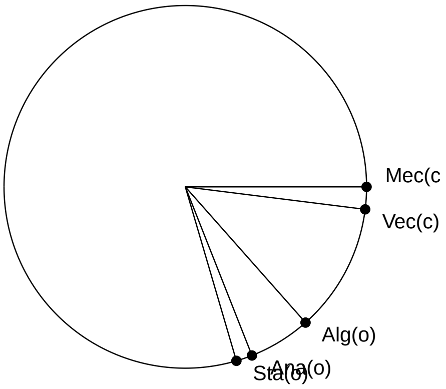

Figure 6: Correlogram of the matrix of student grades on 5 exams.

The approximation this gives to the correlation matrix is given by

```
> angles <- fit_angles(R)
> Rhatcor <- angleToR(angles)
```

and this approximation is shown below in Table 4.

	Mec(c)	Vec(c)	Alg(o)	Ana(o)	Sta(o)
Mec(c)	1.000	0.992	0.663	0.368	0.282
Vec(c)	0.992	1.000	0.751	0.480	0.399
Alg(o)	0.663	0.751	1.000	0.940	0.905
Ana(o)	0.368	0.480	0.940	1.000	0.996
Sta(o)	0.282	0.399	0.905	0.996	1.000

Table 4: Approximation to the correlation matrix obtained by a correlogram.

The RMSE of this approximation is 0.224. In the correlogram in Figure 6 correlations are approximated by cosines of angles. A correlogram with correlations that are linear in the angle is shown in Figure 7 below.

```
> correlogram(R,ifun="lincos",labs=colnames(R),main="",
+             xlim=c(-1.3,1.3),ylim=c(-1.3,1.3))
```

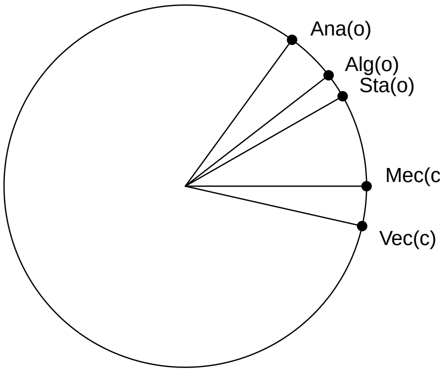

Figure 7: Correlogram with linear interpretation rule of the matrix of student grades on 5 exams.

The approximation to the correlation matrix by using this linear interpretation function is calculated by

```
> theta_lin <- fit_angles(R)
> Rhatcorlin <- angleToR(theta_lin,ifun="lincos")
```

and is shown in Table 5 below.

	Mec(c)	Vec(c)	Alg(o)	Ana(o)	Sta(o)
Mec(c)	1.000	0.921	1.538	0.240	0.182
Vec(c)	0.921	1.000	1.459	0.319	0.261
Alg(o)	1.538	1.459	1.000	0.778	0.720
Ana(o)	0.240	0.319	0.778	1.000	0.942
Sta(o)	0.182	0.261	0.720	0.942	1.000

Table 5: Approximation to the correlation matrix obtained by a linear correlogram.

The RMSE of this last approximation is 0.136.

Acknowledgements

This work was partially supported by grants ECO2011-28875 and CODARSS MTM2012-33236 of the Spanish Ministry of Education and Science. I thank Albert Satorra for supplying the MATLAB code for principal factor analysis. This document was generated by Sweave (Leisch, 2002).

References

Leisch, F. 2002. Sweave: Dynamic generation of statistical reports using literate data analysis. In *Compstat 2002, Proceedings in Computational Statistics*, pages 575–580. Physica Verlag, Heidelberg. URL http:/www.ci.tuwien.ac.at/ leisch/Sweave.

Mardia, K. V., Kent, J. T., and Bibby, J. M. 1979. *Multivariate Analysis.* Academic Press London.

R Core Team 2013. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Satorra, A. and Neudecker, H. 1998. Least-squares approximation of off-diagonal elements of a variance matrix in the context of factor analysis. *econometric theory*, 14(1):156–157.

Trosset, M. W. 2005. Visualizing correlation. *Journal of Computational and Graphical Statistics*, 14(1):1–19.

Weisberg, S. 2005. *Applied Linear Regression.* John Wiley & Sons, New Jersey, third edition.