PERGAMON

Computers & Geosciences 27 (2001) 229–235

COMPUTERS & GEOSCIENCES

Short Note

A program for computing RQ-mode principal components analysis for S-PLUS and R[☆]

E.C. Grunsky*

Alberta Geological Survey, 4th Floor, 4999-98th Avenue, Edmonton, Alta., Canada T6B 2X3

Received 30 September 1999; accepted 25 April 2000

1. Introduction

Principal components analysis (PCA) is a useful tool for evaluating multivariate data when there is a requirement to look for linear relationships in the data that might otherwise be difficult or too time consuming to detect by assessing variables individually or pair-wise. When used as an exploratory tool, patterns may emerge from the data that can be related to geological processes that are manifested within the data. The distinct advantage of applying methods such as principal components analysis is that it reduces the number of variates required to describe the variability in the data.

Principal components analysis summarizes the relationships between m variables using linear combinations of them derived from a symmetric similarity matrix (i.e., variance/covariance matrix, or correlation matrix) or some other suitable metric that describes the variable inter-relationships contained in the observations. The analysis derives orthogonal m eigenvectors and eigenvalues from a symmetric $m \times m$ matrix which define linear combinations of the variables that are weighted according to their significance to the identified sources of variability in the data.

Traditionally, PCA was applied as an R-mode technique for the assessment of the relationships between the variables of a sample population. Q-mode methods are based on the assessment of the relationships between the observations. Both methods use different metrics by which associations are derived. Detailed descriptions of R and Q mode methods are provided by Reyment and Jöreskog (1993) and by Davis (1986).

The interpretation of PCA results is generally carried out by calculating the principal component loadings for both R-mode and Q-mode methods. Using conventional scaling methods loadings of the variables are scaled differently to the loadings of the samples and thus the relationships between samples and variables are difficult to evaluate. Gabriel (1971) developed a scaling method that permits the plotting of R-mode loadings on the same scale as the Q-mode loadings and this technique, which greatly facilitates interpretation, is frequently employed.

The introduction of correspondence analysis (Benzecri, 1970) offered the apparent advantage that a simultaneous R- and Q-mode analysis of multivariate data could be carried out using the same metric. Both R- and Q-mode loadings could be plotted on the same set of diagrams thus enhancing the interpretation of the relationships between observations and variables. The method was introduced to geoscientists by Teil (1975) and later by David et al. (1977) and applied to geochemical data for the purposes of extracting information on the nature of geochemical processes reflected in the data. The appropriateness of this method was challenged by Miesch (1980) and others due to the fact that correspondence analysis was originally developed for categorical data with a metric that reflects a contingency table of probabilities between the categories.

Zhou et al. (1983) studied the use of several metrics and scaling procedures for principal components analysis and developed the RQ-mode method of principal components analysis based on a simultaneous scaling of the relationships between the variables and the observations. A FORTRAN program for RQ-mode PCA, using robust data estimates was published by Zhou (1989).

The development of RQ-mode PCA was based on the premise that most geological, and in particular, geo-

[☆]Code available from http://www.iamg.org/CGEditor/index.htm.

*Tel.: +1-780-422-2454; fax: +1-780-422-1459.

E-mail address: egrunsky@iamg.org (E.C. Grunsky).

PII: S0098-3004(00)00111-4

chemical data, were best measured using the Euclidean distance (or Mahalanobis distance for multivariate data). Many statistical tools have been developed to assist in the interpretation using this metric of statistical measurement.

Given a data matrix of m variables and n observations, a data matrix X can be scaled (i.e. correlation or covariance) to produce a $m{\times}n$ matrix W defined as

$$W = V\Lambda^{1/2}U'$$

where Λ is the diagonal matrix of eigvenalues, V the eigenvector matrix of $n{\times}m$ WW', and U the eigenvector matrix of $m{\times}m$ $W'W$.

By use of the Eckhart–Young theorem (Reyment and Jöreskog, 1993), W can be re-written as

$$W = F^{\mathrm{R}}A^{\mathrm{R}} \quad \text{(R-mode solution)}$$

where $F^{\mathrm{R}} = V$ and $A^{\mathrm{R}} = \Lambda^{1/2}U'$, or

$$W = A^{\mathrm{Q}}F^{\mathrm{Q}} \quad \text{(Q-mode solution)}$$

where $A^{\mathrm{Q}} = V\Lambda^{1/2}$ and $F^{\mathrm{Q}} = U'$. F^{R} and F^{Q} represent the factor loadings for both the R- and Q-mode solutions, and A^{R} and A^{Q} represent the coordinates of the variables and objects (the scores) in the same factor space and can be plotted on the same figures. W is scaled to permit the projection of both F^{R} and F^{Q} in the same coordinate space.

W can be standardized by

$$W_{ij} = (1/n^{1/2})\,(x_{ij} - \bar{x}_j),$$

where $\bar{x}_j = 1/n \sum x_{ij} \quad (i = 1,\, n)$

which yields a variance–covariance matrix from the minor product matrix $W'W$.

W can also be standardized by

$$W_{ij} = (s_j n^{1/2})^{-1}\,(x_{ij} - \bar{x}_j),$$

where $s_j = \left[(1/n)\sum(x_{ij} - \bar{x}_j)^2\right]^{1/2} \quad (i = 1,\, n),$

which results in a correlation matrix from the minor product matrix $W'W$.

The advantage of plotting both the scores of the variables and objects on the same diagram is that the relationships between the two can be more clearly observed. Samples with relative abundance of one variable over another will plot near the location of the score for that variable.

The program presented in this note (program *rpaca.prg*) is written for both the S-PLUS and R programming environments and takes advantage of some of the statistical tools that exist within these packages to further enhance the application of the method.

Of equal importance in the application of multivariate methods is the interpretation and visualization of the results. Both the R and S-PLUS software packages provide tools for plotting the results of PCA. This note includes programs for plotting the variable scores and sample scores on the same set of diagrams in a manner similar to the popular biplot method of Gabriel (1971). Other useful plots for displaying the results of principal components analysis include the "screeplot", a plot of the decreasing ordered eigenvalues, and plots of samples and variable scores projected onto selected principal component axes. Program *rqpca.r.plt* provides the program code for plotting in R, and program *rqpca.s.plt* provides the program code for plotting in S-PLUS.

The FORTRAN program written by Zhou (1989) provides three methods for the robust estimation of the data. Robust estimation methods reduce the influence of outliers when calculating the covariance or correlations of data. Both the R and S software packages offer several robust estimation techniques, which can be built into the routines if required.

2. The S-PLUS and R software packages

S-PLUS is a programming and statistical modelling language that was derived from the S Language developed at AT&T (Becker and Chambers, 1984; Becker et al., 1988). Burns (1998) provides comprehensive coverage of the functionality and use of the S language. Venables and Ripley (1997) is a practical guide for applying S-PLUS to problems in applied statistics. The software environment evolved from a command-line-driven window on UNIX workstations (currently S-PLUS 5) to a fully developed GUI menu system that is currently available for Microsoft Windows operating systems (S-PLUS 2000). The S-PLUS environment offers a large set of statistical and mathematical functions with sophisticated graphical routines that let users visualize the results of data and statistical analyses. S-PLUS was originally developed by StatSci in Seattle, Washington, USA, now a subsidiary of MathSoft Inc.

R, also known as 'GNU S', is a system for statistical computation and graphics. It consists of a language plus a run-time environment with graphics, a debugger, access to certain system functions, and the ability to run programs stored in script files. The language was initially described by Ihaka and Gentleman (1996) and was the result of an effort to combine useful features of two computing languages, S and Scheme. The name R was chosen partly due to the influence of S and partly due to the authors own efforts. R implements a language similar to the S language. The advantage of R is that it is available at no charge. Currently, it does not have as extensive a library of statistical functions as that which exists in S-PLUS. However, there are constant updates to the software with many new functions and libraries being added on a regular basis. Information about R can be obtained from CRAN (1999). Version 0.90.0 of R has been used for the examples shown here.

3. Rqpca program and test data availability

The data used in the following example are those employed by Zhou et al. (1983). They consist of four measurements (airborne radiometric — AREO, Uranium content in ppm — U, Thorium content in ppm — Th and Potassium in percent — K) from 22 sites in the Berea quartz monzonite (Virginia, USA).

Both the data and the program code are available from the Computers & Geosciences ftp site.

4. The rqpca program

The program, *rqpca*, provided with this note is written as an S-PLUS/R function. The program creates a number of attributes that comprise the object created from the execution of the function. In the following example, the object *qzmn.pca.r* is assigned the results of the application of the function *rqpca.*

The function is executed by an assignment of an object to the function with specific arguments.

```
qzmn.pca.r <-rqpca(as.matrix(qzmn),“r”)
```

The arguments of *rqpca* are composed of a data matrix X (in this example qzmn) and a scaling option for use of a correlation or covariance matrix. Both the S-PLUS and R environments require that the data matrix be specified as a numeric matrix. The choice of a correlation or covariance matrix for scaling is determined by the selection of

“r” for correlation matrix, or
“c” for covariance matrix.

On output *qzmn.pca.r* has the following attributes:

`w` — the scaled data matrix,
`eigenvalues` — the eigenvalues,
`eigenvectors` — the normalized eigenvectors,
`eigencontrib` — normalized eigenvalues for a measure of relative significance,
`rscore` — the scores [A^R] of the variables,
`rcr` — the relative contributions of the scores of the variables, and
`rqscore` — the scores [A^Q] of the objects.

These attributes can be described by the following command:

```
attributes(qzmn.pca.r)
$names:
[1]    ‘‘w’’          ‘‘eigenvalues’’       ‘‘eigenvectors’’    ‘‘eigencontrib’’
[2]    ‘‘rscore’’     ‘‘rcr’’               ‘‘rqscore’’
```

The results can be plotted using the following commands to produce a screeplot of the ordered eigenvalues:

```
#Create the Screeplot in S-PLUS
graphsheet()
eigsum <-sum(qzmn.pca.c$b$values)
plot((1:length(qzmn.pca.c$b$values)),qzmn.pca.c$b$values,
xlab=‘‘Ordered Eigenvalue’’,ylab=‘‘Eigenvalue’’,type=‘‘b’’,lab=c(10,10,7)
,mgp=c(2,1,0),main=‘‘Screeplot of RQ-Mode PCA (Covariance Matrix)’’)
```

These commands employ the syntax used in plotting functions based on the command line interface.

To plot the first and second principal components for both the R-mode and Q-mode scores onto the principal component axes, the following command is used:

```
#Create a plot of PC1 vs. PC2 in S-PLUS
graphsheet()
rqpca.s.plt(qzmn.pca.r,qzmn,1,2,T,F,‘‘’’)
mtext(side=3,‘‘RQPCA (Covariance Matrix)’’,cex=1,line=−2,outer=T)
mtext(side=3,‘‘QZMN Data’’,cex=1,line=−1,outer=T)
```

The components (1 and 2) are identified as “, 1, 2, T,...” in the plotting function call parameters.

Examples of the results of the computation and graphics for both the S-PLUS and R environments follows.

5. Example of the programs in S-PLUS

A sample session for RQ-mode PCA using S-PLUS 2000 (Mathsoft, 1999), using one of a variety of import options (ODBC database, EXCEL spreadsheet) or the *scan* or *read.table* function to read the Berea data into the S-PLUS environment, is as follows:

```
>qzmn<-read.table(file=‘‘d:/qzmn.dat’’,header=T,sep=‘‘\t’’)
>qzmn
        Aero            U               Th              K
 1       240            0.63             2.05           0.13
 2       360            2.18             5.31           0.31
 3       420            2.26             5.61           0.34
 4       500            1.71             6.44           0.70
 5       580            2.38             7.99           1.73
 6       700            3.83             8.32           4.26
 7       600            3.79             9.46           1.53
 8       650            4.09            14.71           3.11
 9       770            4.21            12.00           1.90
10       930            4.72            12.78           2.92
11      1020            6.24            16.31           2.29
12      1000            5.24            14.51           1.88
13      1000            4.73            15.79           4.64
14      1040            4.67            10.30           4.17
15      1150            5.08            13.11           3.97
16      1000            5.27            13.40           4.36
17       960            5.61            10.31           2.05
18       420            2.33             6.83           0.47
19       370            2.64             9.88           0.58
20       400            2.29             6.02           0.34
21       480            2.32             6.14           0.32
22       730            5.94            12.86           1.35
```

```
> #RQ-mode analysis using the correlation matrix (‘‘r’’ option)
>qzmn.pca.r<− rqpca (as.matrix(qzmn), ‘‘r’’)
[1]  ‘‘Eigenvalues’’
[1]  3.23813273     0.37283895   0.14789201   0.05931812
qzmn.pca.r
$w:
     Aero         U             Th            K
 1   −0.35088289  −0.421778983  −0.432975848  −0.254659520
 2   −0.25861886  −0.211198261  −0.255560389  −0.229753328
 3   −0.21248685  −0.200329578  −0.239233813  −0.225602296
 4   −0.15097750  −0.275051770  −0.194063619  −0.175789913
 5   −0.08946815  −0.184026555  −0.109709643  −0.033271150
 6    0.00279588   0.012968314  −0.091750410   0.316799210
 7   −0.07409081   0.007533973  −0.029709421  −0.060944696
 8   −0.03564747   0.048291532   0.256005659   0.157676319
 9    0.05661656   0.064594556   0.108522256  −0.009748636
10    0.17963527   0.133882406   0.150971354   0.131386450
11    0.24883329   0.340387372   0.343080731   0.044214780
12    0.23345595   0.204528842   0.245121275  −0.012515990
13    0.23345595   0.135240992   0.314781333   0.369378948
14    0.26421062   0.127089480   0.016004992   0.304346114
```

```
15    0.34878598    0.182791477    0.168930587    0.276672568
16    0.23345595    0.208604598    0.184712944    0.330635983
17    0.20270127    0.254796498    0.016549211    0.011006524
18   −0.21248685   −0.190819481   −0.172839071   −0.207614491
19   −0.25093020   −0.148703337   −0.006852214   −0.192394041
20   −0.22786419   −0.196253823   −0.216920826   −0.225602296
21   −0.16635484   −0.192178067   −0.210390195   −0.228369651
22    0.02586189    0.299629813    0.155325107   −0.085850888

$eigenvalues:
[1]    3.23813273     0.37283895     0.14789201     0.05931812

$eigenvectors:
     [,1]           [,2]           [,3]           [,4]
[1,] 0.5218121      0.0793351      0.5592382      0.6392735
[2,] 0.5102128     −0.4276586      0.3424457     −0.6629645
[3,] 0.5016891     −0.3961704     −0.7201665      0.2696614
[4,] 0.4644384      0.8086180     −0.2265916     −0.2812297

$eigencontrib:
[1]    84.808238      9.764830       3.873362       1.553570

$rscore:
     [,1]           [,2]           [,3]           [,4]
[1,] 0.9389911      0.04844244     0.21506474     0.15569706
[2,] 0.9181184     −0.26113064     0.13169343    −0.16146707
[3,] 0.9027801     −0.24190379    −0.27695249     0.06567688
[4,] 0.8357482      0.49374652    −0.08713971    −0.06849437

$rcr:
     [,1]           [,2]           [,3]           [,4]
[1,] 92.36903       0.2458416     4.8455358      2.5395936
[2,] 88.30815       7.1436317     1.8169024      2.7313120
[3,] 85.38220       6.1303988     8.0355192      0.4518855
[4,] 73.17357      25.5394465     0.7954917      0.4914883

$rqscore:
     [,1]           [,2]           [,3]           [,4]
 1   −0.7337849     0.118150018    0.0288548709   0.010175283
 2   −0.4776246    −0.014734020    0.0191527047  −0.029612563
 3   −0.4378883    −0.018833724    0.0359749820  −0.004091884
 4   −0.3981200     0.040385810    0.0009679414   0.082939513
 5   −0.2110708     0.088162635   −0.0265049417   0.044580830
 6    0.1091791     0.287194144    0.0002960457  −0.120645098
 7   −0.0780275    −0.046610956   −0.0036491519  −0.043231075
 8    0.2077039     0.002597661   −0.2234930228  −0.030112466
 9    0.1124170    −0.074008964   −0.0221628598   0.025375445
10    0.2988060     0.003426489    0.0078107897   0.029838165
11    0.4961691    −0.225993935   −0.0013726628   0.013488921
12    0.3433351    −0.176177714    0.0269054034   0.083266469
13    0.5202980     0.134663700   −0.1335229405   0.040586280
14    0.3520905     0.206369314    0.1107894044   0.003371803
15    0.4885113     0.106295748    0.0733007899   0.069530903
16    0.4744814     0.123490112   −0.0059501614  −0.032229941
17    0.2491868    −0.090540818    0.1862000970  −0.037972151
18   −0.3913724    −0.034659168   −0.0126594805   0.002448655
```

```
19   −0.2996016      −0.109171956   −0.1427229747  −0.009568856
20   −0.4326387      −0.030636464    0.0127020460  −0.010607338
21   −0.3964718      −0.032324620    0.0444200699   0.028551202
22    0.2044225      −0.257043292    0.0246630509  −0.116082097
```

The graphical presentation of the significance of the eigenvalues can be shown in a screeplot (Fig. 1) using the commands:

```
#Create the Screeplot in S-PLUS
graphsheet()
plot((1:length(qzmn.pca.r$eigenvalues)),qzmn.pca.r$eigenvalues,
xlab=‘‘Ordered Eigenvalue’’,ylab=‘‘Eigenvalue’’,type=‘‘b’’,lab=c(10,10,7)
,mgp=c(2,1,0),main=‘‘Screeplot of RQ-Mode PCA (Correlation Matrix)’’)
```

A plot of the first two principal components is generated by the following commands using the function *rqpca.s.plt* to plot (Fig. 2) both the sample points (observations) and variables plot on the same diagram:

```
#Create a plot of PC1 vs. PC2 in S-PLUS
graphsheet()
rqpca.s.plt(qzmn.pca.r,qzmn,1,2,T,F,‘‘ ’’)
mtext(side=3,‘‘RQPCA (Correlation Matrix)’’,cex=1,line=−2,outer=T)
mtext(side=3,‘‘QZMN Data’’,cex=1,line=-1,outer=T)
```

6. Example of the programs in R

A sample session of the use of RQ-mode PCA using R follows. As R uses functions that are not available in S-PLUS there are variations on how the RQPCA procedures are executed. A summary of the procedures is listed below.

Using import routines such as *scan* or *read.table* in R, the quartz monzonite data is read into the R environment.

```
qzmn<-read.table(file=‘‘e:/qzmn.dat’’,header=T,sep=‘‘\t’’)

#RQ-mode analysis using the correlation matrix


qzmn.pca.r<-rqpca(as.matrix(qzmn),‘‘r’’)
[1]  ‘‘Eigenvalues’’
[1]  3.23813273       0.37283895      0.14789201      0.05931812
>
>#RQ-mode analysis using the covariance matrix
>qzmn.pca.c<-rqpca(as.matrix(qzmn),‘‘c’’)
[1]  ‘‘Eigenvalues’’
[1]  7.340920e+04     4.874014e+00    8.017630e-01    2.547302e-01
>
```

```
#Create the Screeplot in R
win.graph(width=9,height=6)
plot((1:length(qzmn.pca.r$eigenvalues)),qzmn.pca.r$eigenvalues,
xlab=‘‘Ordered Eigenvalue’’,ylab=‘‘Eigenvalue’’,type=‘‘b’’,lab=c(10,10,7)
,mgp=c(2,1,0),main=‘‘Screeplot of Eigenvalues RQ-Mode PCA (Correlation Matrix ’’)
```

```
#Create a plot of PC1 vs. PC2 in R
win.graph(width=9,height=6)
rqpca.r.plt(qzmn.pca.r,qzmn,1,2,T,F,‘‘ ’’)
mtext(side=3,‘‘RQPCA (Correlation Matrix)’’,cex=1,line=-2,outer=T)
mtext(side=3,‘‘QZMN Data’’,cex=1,line=-1,outer=T)
```

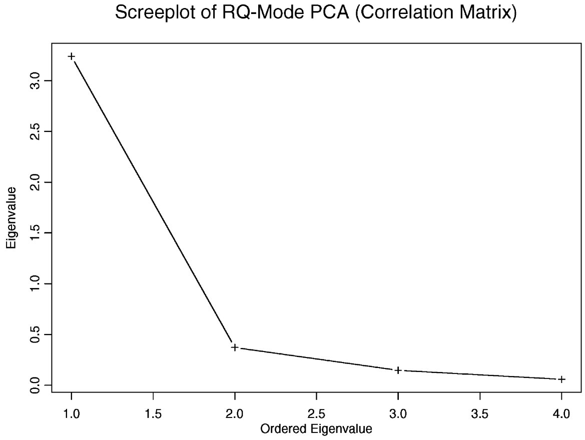

Fig. 1. Screeplot of eigenvalues from QZMN data using S-PLUS *rqpca* functions.

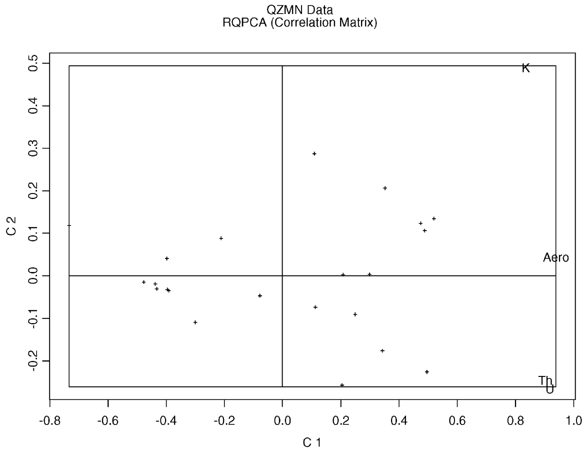

Fig. 2. Plot of first two principal components based on correlation matrix for QZMN data. Note that association (or lack thereof) between samples and variables can be visually assessed in this type of diagram.

Acknowledgements

The author gratefully acknowledges helpful comments to this manuscript provided by Bob Garrett, Geological Survey of Canada and an anonymous reviewer.

References

Becker, R.A., Chambers, J.M., 1984. S: An Interactive Environment for Data Analysis and Graphics, Wadsworth, Belmont, CA, 550pp.

Becker, R.A., Chambers, J.M., Wilks, A.R., 1988. The New S Language, A Programming Environment for Data Analysis and Graphics, Wadsworth & Brooks/Cole, Advanced Books and Software, 702pp.

Benzecri, J.P., 1970. La practique de l'analyse des correspondence. Cahier no. 2, Laboratoire de Statistique Mathematique, Faculte de Sciences, Paris, 35pp.

Burns, P.J., 1998. S Poetry, http://www.seanet.com/~pburns/Spoetry/

CRAN, 1999. The Comprehensive R Network, http://www.ci.tuwien.ac.at/R/contents.html

David, M., Dagbert, M., Beauchmin, Y., 1977. Statistical analysis in geology: correspondence analysis method. Quarterly of the Colorado School of Mines 60 (1), 1–60.

Davis, J.C., 1986. Statistics and Data Analysis in Geology, 2nd edn. Wiley, New York, 646pp.

Gabriel, K.R., 1971. The biplot graphic display of matrices with application to principal components analysis. Biometrika 58 (3), 453–467.

Ihaka, R., Gentleman, R., 1996. R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics 5 (3), 299–314.

Mathsoft, 1999. SPLUS 2000, Programmers Guide. Data Analysis Products Division, Mathsoft Inc., Seattle, WA, USA, 868pp.

Miesch, A.T., 1980. Scaling variables and interpretation of eigenvalues in principal components analysis of geological data. Mathematical Geology 12 (6), 523–538.

Reyment, R.A., Jöreskog, K.G., 1993. Applied Factor Analysis in the Natural Sciences. Cambridge University Press, New York, 371pp.

Teil, H., 1975. Correspondence factor analysis: an outline of its method. Mathematical Geology 7 (1), 3–12.

Venables, W.N., Ripley, B.D., 1997. Modern Applied Statistics with S-PLUS, 2nd edn. Springer, New York, 548pp.

Zhou, D., 1989. ROPCA A Fortran program for robust principal components analysis. Computers & Geosciences 15 (1), 59–78.

Zhou, D., Chang, T., Davis, J.C., 1983. Dual extraction of R-mode and Q-mode factor solutions. Mathematical Geology 15 (5), 581–606.