ViSta

The Visual Statistics System

Principal Components Analysis

Forrest W. Young & Pedro Valero

Visual Statistics Project
Report Number 99-1
October, 1999

THE L.L. THURSTONE
PSYCHOMETRIC LABORATORY
UNIVERSITY OF NORTH CAROLINA

CHAPEL HILL, N.C.
U.S.A. 27599-3270

Principal Components Analysis

Forrest Young and Pedro Valero

This chapter presents ViSta-PrnCmp, the module for Principal Components Analysis (PCA) in ViSta. This procedure is capable of analyzing numerical variables so they can be represented on a lower dimensionality space. The visualization for ViSta-PrnCmp includes a scatterplot-matrix of component scores; a bi-dimensional biplot; a tri-dimensional (spin-plot) version of the biplot; a box-diamond-dot plot and a scree plot of the relative fit of the components. The report includes the eigenvalues and eigenvectors of the correlation or covariance matrix; the correlations of the variables with the dimensions and other optional output. ViSta-PrnCmp also allows creating data objects such as component scores that can be submitted for further analysis.

1 Introduction

Principal components analysis (PCA) is a statistical procedure that allows the researcher to find a reduced number of dimensions that account for the maximum possible amount of variance in the data matrix. This is useful for exploratory analysis of multivariate data because the new dimensions, called principal components, can be represented graphically and admit a more succinct interpretation than the original data matrix.

The type of data appropriate for analysis with ViSta-PrnCmp will be described now. PCA only assumes that the data is numerical and measured at interval or ratio levels. This analysis does not involve a hypothesis testing procedure, so it is not necessary to assume that the data sampled comes from a normally distributed population or under any other theoretical probability distribution. However, if you analyse data composed by variables asymmetrically distributed, with heteroskedastic variances or that keep non-linear relationships among them, the results may be distorted by the influence of a few observations and therefore their interpretation will be uncertain. Hence, prior to carrying out PCA, examination of the visualization for numerical data is recommended.

2 Example

What follows is an example based on data corresponding to the rate, per 100,000 population, of each of the seven major crime classifications, in each of the 50 USA states in 1980. This data can be found in the file Crime.lsp. The visualization (the spreadplot for multivariate numerical data) is shown in figure 1.

The connected Box Plot may be used to check whether boxplots appear symmetric for all the variables. Clicking on the variables window is useful for confirming this impression by way of the normal probability and the frequency plots that arise automatically. Also, the scatterplot/spin-plot will provide evidence on the linearity of the relationships among variables. Furthermore, it should be examined to see if there are obvious outliers in the data.

We carried out the Principal Component Analysis with the data as provided, ignoring the slight asymmetry of the Auto-Theft variable (which can be seen in the boxplot, normal-probability plot and frequency plot), and being wary of the two possible outliers (seen in the boxplot). We choose to analyze the correlations among the variables, even though the units of the variables are the same (crime/100,000 population) since, if the covariance matrix had been used, those variables with more variance would have unduly influenced the analysis. The visualizations and report shown in figures 2 and following present the result of the PCA analysis provided by ViSta.

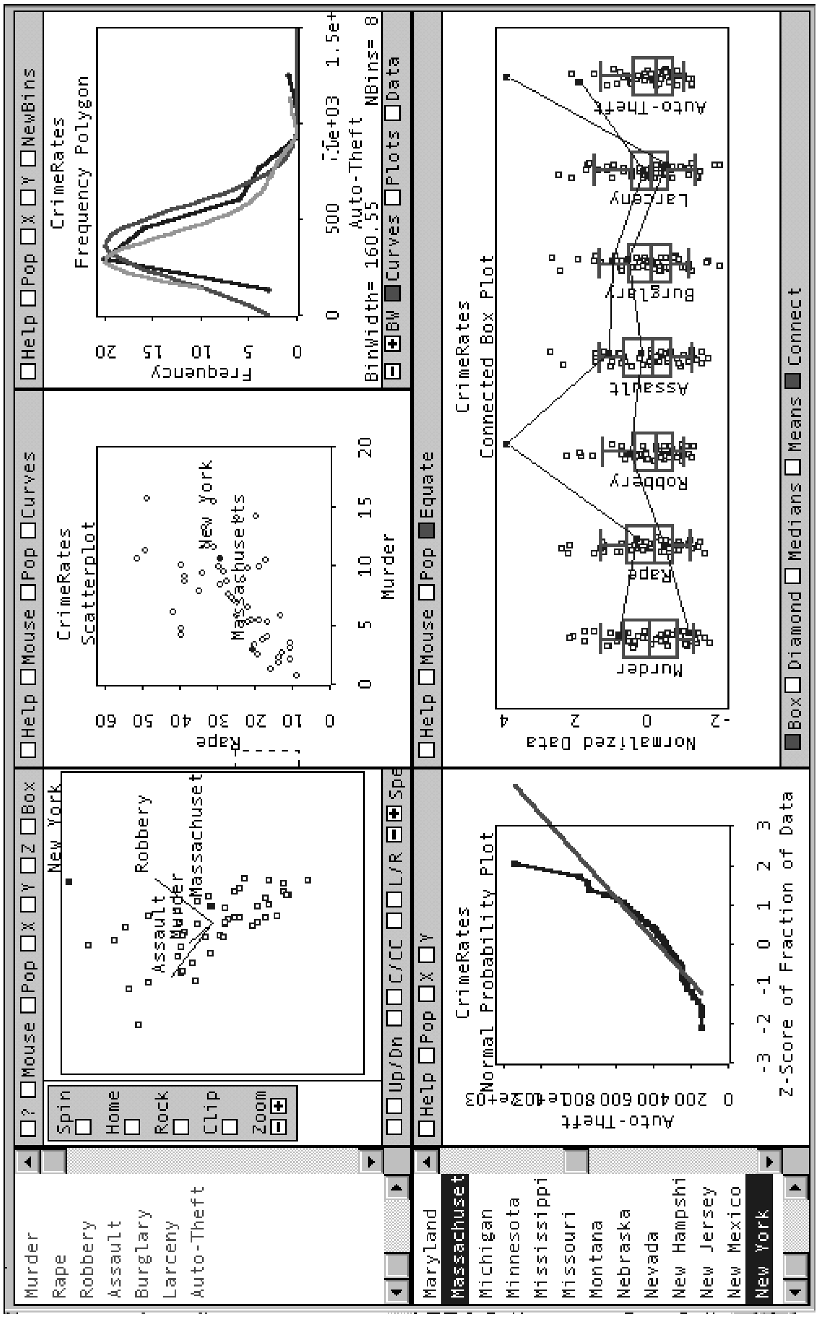


Figure 1: Multivariate Visualization for the Crime Data

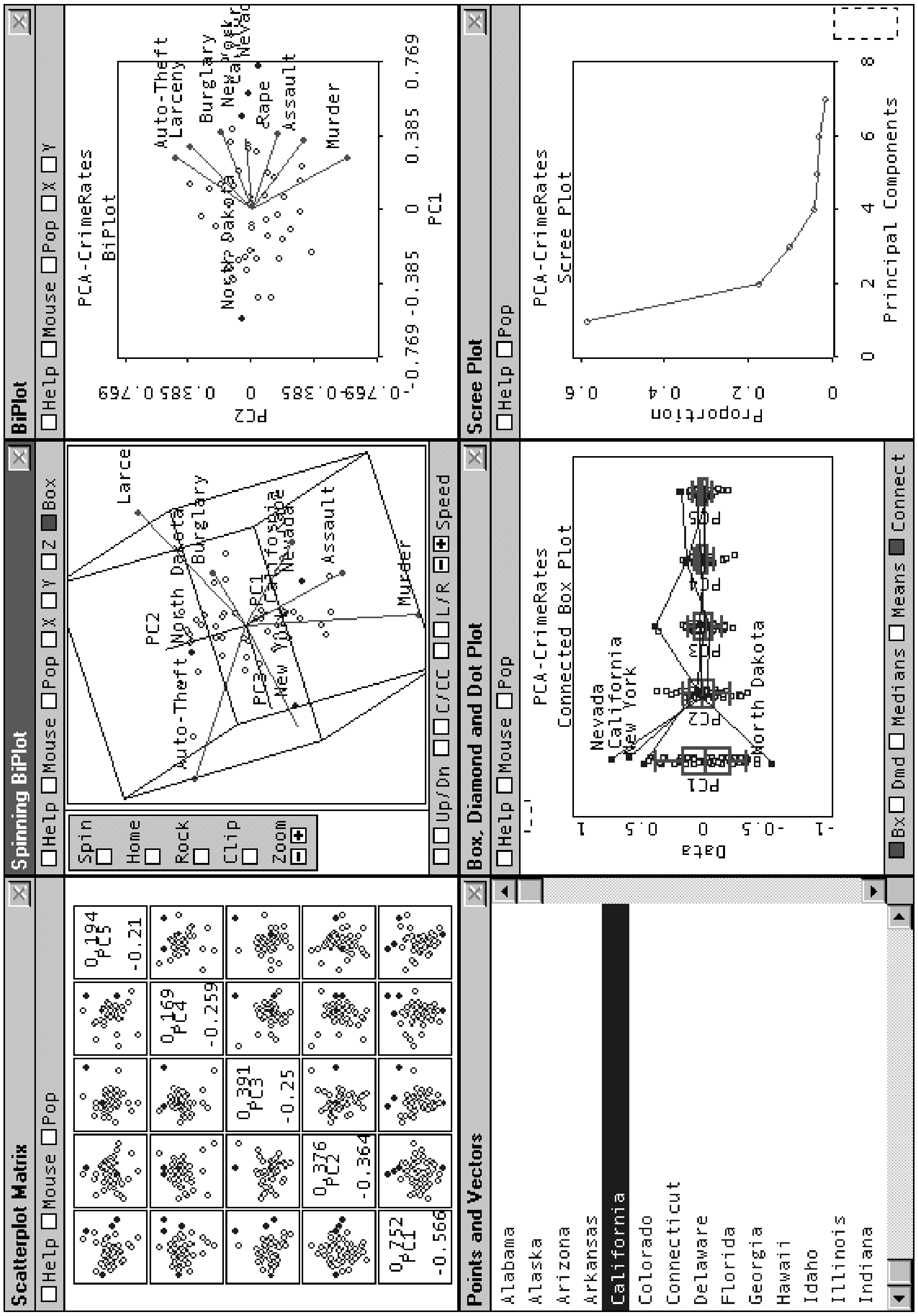


Figure 2: Principal Components Visualization for the Crime Data

3 Visualization

.The principal components visualisation has five plots. Clicking on the scatterplots in the Scatterplot Matrix window changes the principal components shown in the Spinning Biplot window, the Biplot window and the Box, Diamond and Dot Plot window. There is a limit of five in the number of components shown in these plots. Each point in these plots represents the value of an observation in the corresponding principal component. The labels of the observations (blue) can be examined in the Points and Vectors window. This window also shows the labels of the variables analyzed (red). Finally, the Scree Plot shows the proportion of variance explained by each of the principal components.

Biplot: A Biplot is an enhanced scatterplot that uses both points and vectors to represent structure. As used in PCA, the axes of a biplot are a pair of principal components. A biplot uses points to represent the scores of the observations on the principal components, and uses vectors to represent the coefficients of the variables on the principal components.

The relative location of the points can be interpreted. Points that are close together correspond to observations that have similar scores on the components displayed in the plot. When these components fit the data well, the points also correspond to observations that have similar values on the variables.

Both the direction and length of the vectors can be interpreted. Vectors point away from the origin in some direction. Vectors pointing in the same direction correspond to variables that have similar response profiles and can be interpreted as having similar meaning in the context set out by the data. Vectors pointing in opposite directions correspond to variables with similar but reversed response profiles, such as when there are negative correlations. Long vectors are more strongly related to the components being displayed than are short vectors. Long vectors are more important in interpreting the meaning of the components.

We obtained figure 3 by selecting some points and vectors in the biplot. A previous exploration using the brush had led us to observe a cluster of southern states in the principal plane (this cluster does not include Florida, which has a similar crime pattern to northern states). Observing the labels of the vectors we may conclude that southern states have larger proportion of crimes against the people than against the property.

Scree Plot: The Scree plot (figure 4) shows the relative fit of each principal component. It does this by plotting the proportion of the variance of the data that is fit by each component versus the number of components. The plot shows the relative importance of each component in fitting the data. The components will always be sorted according to their relative importance, so initial components will always explain more variance than those placed in subsequent positions.

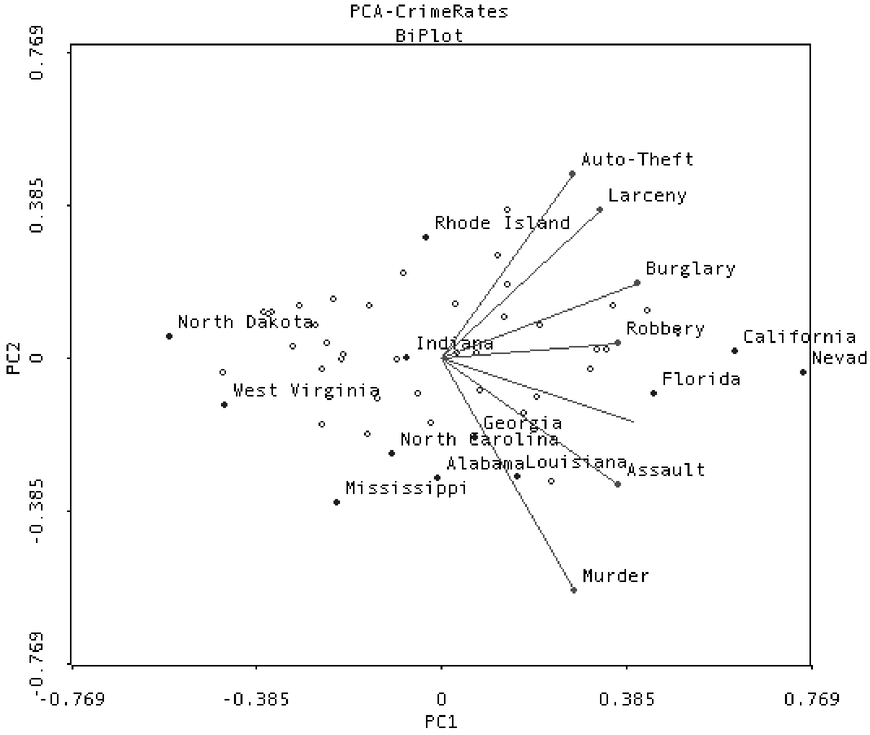


Figure 3: Biplot for the Crime Data

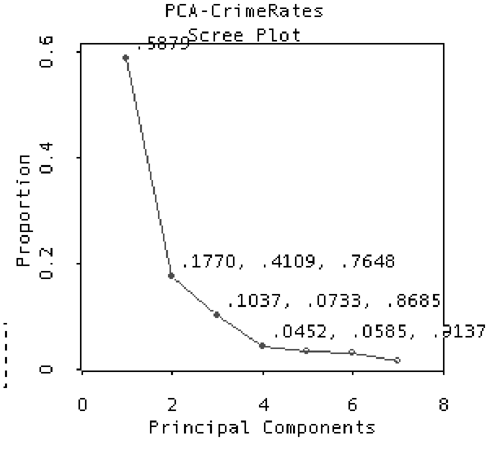


Figure 4:Scree Plot for the Crime Data

The numbers beside the points provide information about the fit of each component. The first number is the proportion of the variance of the data that is accounted for by the component. The second number is the difference in variance from the previous component. The third number is the total proportion of variance accounted for by the component and the preceding components.

The Scree plot can be used to aid in the decision about how many components are useful. It is used to make this decision by looking for an elbow (bend) in the curve. If there is one, then the components following the bend account for relatively little additional variance, and can perhaps be ignored.

The information in the scree plot indicates that the two first components of the Crime data account for the 76.48% of variance of the data. This means that restricting the analysis to the interpretation of the biplot in figure 3 only 23.52% of the variance would not be considered (spread over five components). Actually, the first component in the result accounts for almost half of the variance of the data. It can be interpreted in terms of states with more criminality (right side of the plot) versus those/states with less criminality (left side of the plot) in the different categories. This component could be used for subsequent analysis requiring a measurement of "Criminality" in each state. Study of the projection of the points onto the second dimension suggests it has states with proportionately low property crime on the bottom (southern states) and those with proportionately high property crime at the top (industrialized states).

4 Report

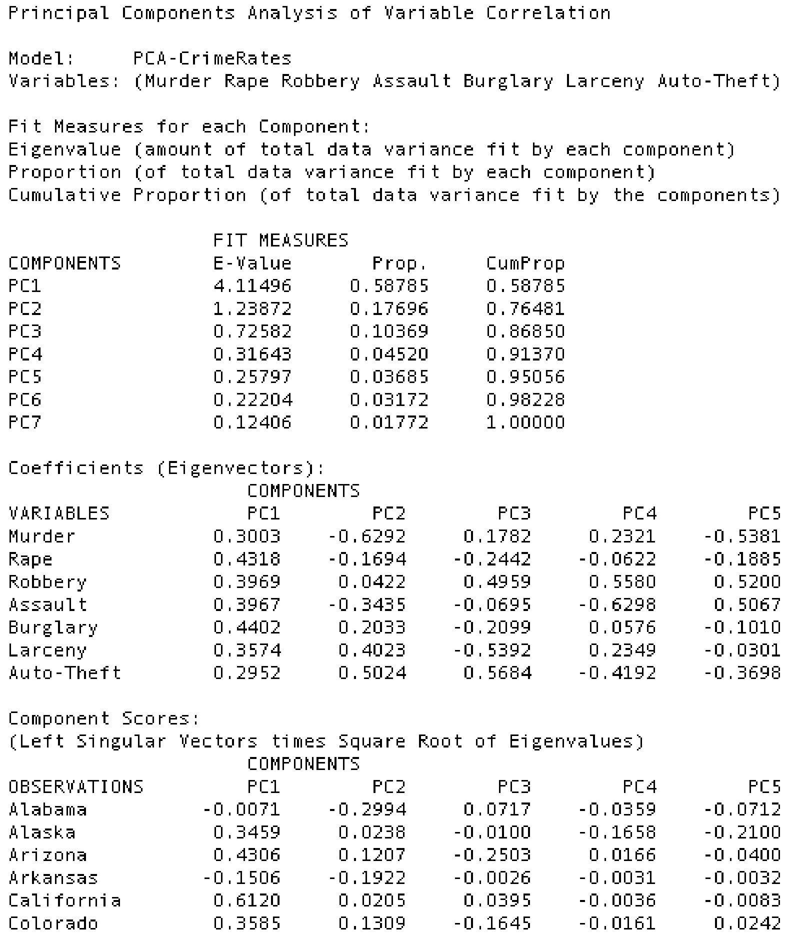
The report is shown in Figure 5 (only a portion of the entire report is shown in the figure). It displays numeric information about the principal components analysis. The top portion shows the measures of fit: eigenvalues, proportion of variance explained and cumulative proportion of variance explained. When the analysis is based on correlations, the sum of the eigenvalues equals the number of variables that were analyzed, and, when it is based on covariances, it equals the sum of the individual variances. The proportion of variance explained is each eigenvalue divided by the sum of the eigenvalues. The second part shows the coefficients or eigenvectors. The eigenvectors for a principal component are the cosines of the angles of the variables with this principal component. This information is used to draw the lines in the biplot in figure 3.

The component scores are the coordinates of the observations in the space of the principal components. They are used to draw the points shown in the biplot in figure 3.

ViSta-PrnCmp can provide additional output (not shown). This output includes the correlation or covariance matrix the coordinates and the contributions.

The correlation matrix is a first step in the computation of the PCA. This procedure will find a reduced number of components explaining the greater part of the variance in the original data only if the variables in the data are substantially correlated.

The output for the coordinates includes the Variable Coordinates also known as loadings or correlations of the variables with the components. These correlations are between -1 and 1 and can be used to give meaning to the components. Positive correlation of a component with a variable imply that observations with high values in the component will have generally high values in that variable. Negative correlations will denote the opposite.



```
Principal Components Analysis of Variable Correlation

Model:     PCA-CrimeRates
Variables: (Murder Rape Robbery Assault Burglary Larceny Auto-Theft)

Fit Measures for each Component:
Eigenvalue (amount of total data variance fit by each component)
Proportion (of total data variance fit by each component)
Cumulative Proportion (of total data variance fit by the components)

                  FIT MEASURES
COMPONENTS        E-Value       Prop.     CumProp
PC1               4.11496     0.58785     0.58785
PC2               1.23872     0.17696     0.76481
PC3               0.72582     0.10369     0.86850
PC4               0.31643     0.04520     0.91370
PC5               0.25797     0.03685     0.95056
PC6               0.22204     0.03172     0.98228
PC7               0.12406     0.01772     1.00000

Coefficients (Eigenvectors):
                     COMPONENTS
VARIABLES            PC1        PC2        PC3        PC4        PC5
Murder            0.3003    -0.6292     0.1782     0.2321    -0.5381
Rape              0.4318    -0.1694    -0.2442    -0.0622    -0.1885
Robbery           0.3969     0.0422     0.4959     0.5580     0.5200
Assault           0.3967    -0.3435    -0.0695    -0.6298     0.5067
Burglary          0.4402     0.2033    -0.2099     0.0576    -0.1010
Larceny           0.3574     0.4023    -0.5392     0.2349    -0.0301
Auto-Theft        0.2952     0.5024     0.5684    -0.4192    -0.3698

Component Scores:
(Left Singular Vectors times Square Root of Eigenvalues)
                     COMPONENTS
OBSERVATIONS         PC1        PC2        PC3        PC4        PC5
Alabama          -0.0071    -0.2994     0.0717    -0.0359    -0.0712
Alaska            0.3459     0.0238    -0.0100    -0.1658    -0.2100
Arizona           0.4306     0.1207    -0.2503     0.0166    -0.0400
Arkansas         -0.1506    -0.1922    -0.0026    -0.0031    -0.0032
California        0.6120     0.0205     0.0395    -0.0036    -0.0083
Colorado          0.3585     0.1309    -0.1645    -0.0161     0.0242
```

Figure 5: Report for the Crime Data

This information is redundant with that on eigenvalues and eigenvectors but many researchers prefer to interpret the correlations. The sum of the squares of the correlations of a principal component with all the variables adds up to the correspondent eigenvalue and the sum of the squares of the correlations of a variable with all the components equal to 1. This last property allows for exploration into the way that the variance of each variable is spread along the components.

The observation coordinates are equal to the component scores multiplied by the root square of the number of observations.

Finally, the information in the absolute and relative contributions of the observations and variables (not shown) can be used for helping the process of interpreting the components. Readers interested can consult Jambu (1989).

5 Using PrnCmp

The Principal Component Analysis procedure works on multivariate data. The computations can not be carried out when there are more variables than cases. No missing data is allowed (but the Impute Missing Data command in the Transformations menu in ViSta can be used to fix this problem)..

5.1 Analysis Options

The only option in ViSta-PrnCmp is to choose between a matrix of correlations or of covariances as input. The option of correlations involves standardizing the variables to have mean equal to zero and variance equal to 1. This is normally the most sensible option. PCA on covariance matrices causes problems when the variables show large differences in variances. The major problem is that variables with large variances will always get larger weights and variables with small variance will always get insignificant weights. If the scales of the different variables are arbitrary, the use of correlation matrices can solve this problem. Covariance matrices should only be used when the scale of the variables is similar.

A principal components analysis can also be performed by typing (principal-components) in the ViSta Listener window. This gives you the default analysis, which is based on the correlation matrix. There are several keywords that can be used additionally when typing in the listener window. These keywords and their default values are given below:

`:covariances` followed by t (analyze covariances) or nil (analyze correlations, the default).

`:data` followed by the name of the data to be analyzed (default: current-data)

`:title` Used to specify a title of the report. The title is entered with quotes. Default is “`Analysis of Variance`”..

`:dialog` The value `t` indicates that the parameter dialog box should be displayed, whereas `nil` indicates that the dialog box should not be displayed. Default is `nil`

For example, for an analysis of principal components based on covariances type (principal-components :covariances t).

5.2 Report

The report for the PCA procedure can be obtained by selecting the Report Model item from the Model menu, by typing (report-model) or by sending the model object the :report message. The Report dialog box shown in figure 7 will then appear. A basic PCA report can be obtained by simply clicking OK. This includes eigenvalues, eigenvectors and scores of observations. Longer results can be obtained by selecting the options shown in the dialog box. This will produce the matrix analyzed, the contributions for the variables (also called correlations with components or loadings) and other additional output.

5.3 Statistical Visualization

You can see the visualization of the analysis results by using the Visualize Model item from the Model menu, by typing (visualize-model)in the listener window, or by sending the model object the :visualize-model message. The visualization is a spreadplot of five interacting plots. Two of these plots, the biplot and the scree plot, have already been explained in the introduction section. This section describes the windows contained in the spreadplot.

The Points and Vectors window shows the labels of the observations (points-blue) and of the variables (vectors-red). Clicking on any of the labels makes that the corresponding point highlights in other plots in the spreadplot.

The Scatterplot-Matrix shows the scores of the observations in the five first principal components. If the mouse is in 'focus on variables' () mode, clicking on any of the small scatterplots will change the variables shown in the biplot, the spinning biplot and the boxplot. Clicking along the diagonal will change the orientation in the spinning plot. This plot also supports the selecting and brushing mouse modes.

The Biplot shows a plot of vectors and points of the data analysed. First and second principal components are shown by default. Other dimensions can be chosen using the Scatterplot-Matrix or the X and Y buttons in the window. The plot supports selecting and brushing mouse modes.

The Spinning biplot shows a 3-D version of the biplot. Standard spin-plots controls, and selecting, brushing and hand rotate ( ) mouse modes are supported.

Finnally, the Scree plot shows a plot of the proportion of variance explained by the principal components. This plot can be brushed to examine interactively this proportion, the cumulative proportion and the actual eigenvalue for each component.

5.4 Create Data

Most of the results in the Report window can be converted into data objects. This allows for further analysis or visualization of them. Figure 9 shows the options in ViSta.

Component Scores are the values of the observations in the principal components computed by ViSta-PrnCmp. If the variance accounted for a few components of the PCA is high, the analysis could go on using them as new variables. For example, the scores in the first principal component for the PCA of the Crime data could be used as a measure of the "Criminality" of the states in USA. Other output can be used for further diagnosis of the results of the PCA.

6 Algorithm

Principal component analysis is straightforward as long as you have a way of computing the eigenvalues and eigenvectors of a covariance matrix. The procedure followed in ViSta is:

1. Normalize data to mean zero and variance 1 and divide the data by the square root of the number of observations minus 1 in case a correlation matrix analysis is requested. Center data to have mean zero computing each variable minus its mean in case a covariance matrix is requested and divide the data by the square root of the number of observations minus 1 in case a covariance matrix analysis is requested.
2. Compute the singular value decomposition of the data. This results in the left and right singular vectors and the singular values of the data. ViSta uses a function in Lisp-Stat to carry out this computation.
3. Computation of component scores. This is obtained by multiplying the matrix of left singular vectors by the singular values. ViSta also computes the observa-

tion coordinates that results of multiplying the component scores by the root square of the number of cases. Observation coordinates and component scores are equivalent, but the second is more compatible with it is obtained in other statistical packages.

4. Computation of eigenvalues. This is computed squaring the singular values obtained in step 2.
5. Computation of coefficients (eigenvectors). This is the matrix of right singular vectors obtained in step 2.

7 References

Jambu, M. (1991). Exploratory and Multivariate data analysis. Academic Press.