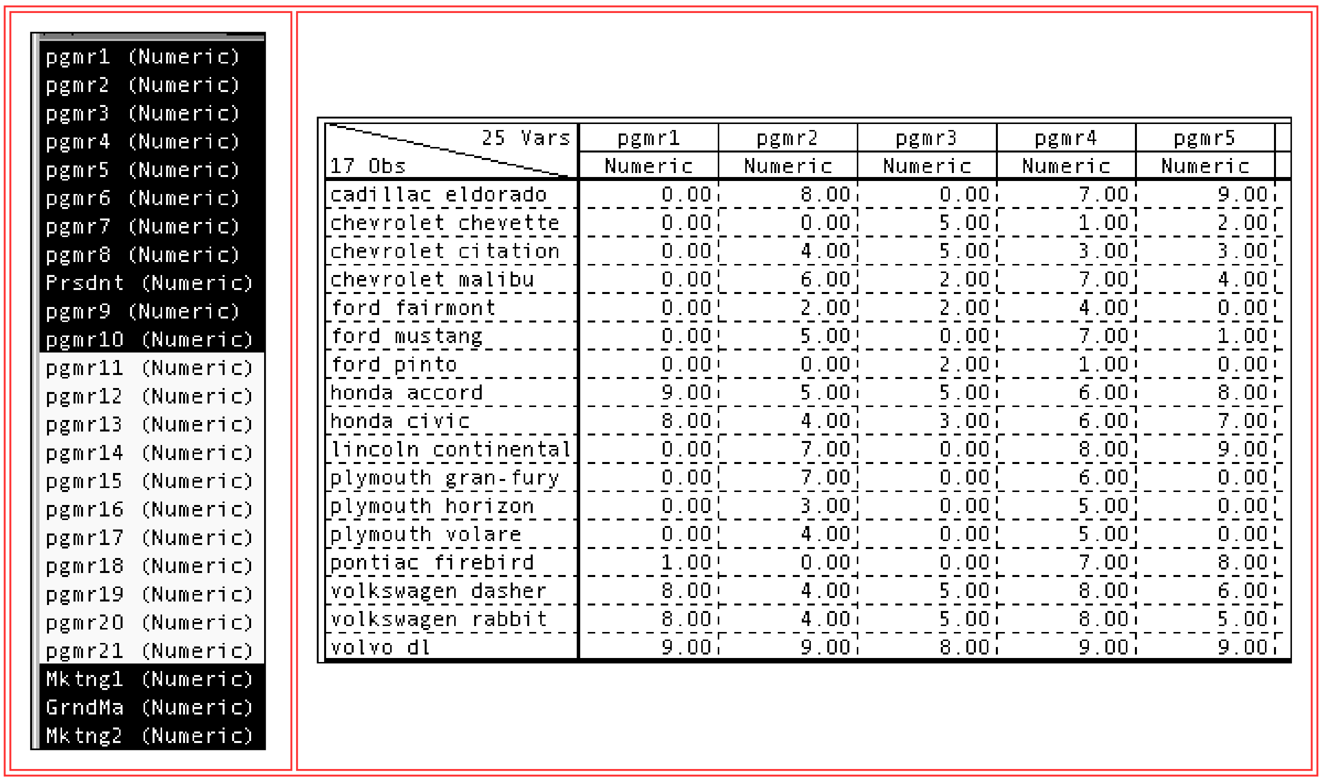
Principal Components:
Car Preference Example

Forrest Young's Notes

Forrest Young and Warren Sarle gathered the following judgments of preference for a set of automobiles from the staff of a large local statistical software house in 1982.

The data are judgments, on a scale of 0-10 of the preference the judge has for the automobile (0 = no preference; 10 = maximum preference). Altogether, there were 25 subjects and 17 Automobiles. Note that for preference data the variables correspond tot he judges and the observations to the Automobiles. Thus, we have more columns than rows.

The Var Window (in this case, subjects) and a portion of the data are shown below:



pgmr1 (Numeric)
pgmr2 (Numeric)
pgmr3 (Numeric)
pgmr4 (Numeric)
pgmr5 (Numeric)
pgmr6 (Numeric)
pgmr7 (Numeric)
pgmr8 (Numeric)
Prsdnt (Numeric)
pgmr9 (Numeric)
pgmr10 (Numeric)
pgmr11 (Numeric)
pgmr12 (Numeric)
pgmr13 (Numeric)
pgmr14 (Numeric)
pgmr15 (Numeric)
pgmr16 (Numeric)
pgmr17 (Numeric)
pgmr18 (Numeric)
pgmr19 (Numeric)
pgmr20 (Numeric)
pgmr21 (Numeric)
Mktng1 (Numeric)
GrndMa (Numeric)
Mktng2 (Numeric)

25 Vars	pgmr1	pgmr2	pgmr3	pgmr4	pgmr5
17 Obs	Numeric	Numeric	Numeric	Numeric	Numeric
cadillac eldorado	0.00	8.00	0.00	7.00	9.00
chevrolet chevette	0.00	0.00	5.00	1.00	2.00
chevrolet citation	0.00	4.00	5.00	3.00	3.00
chevrolet malibu	0.00	6.00	2.00	7.00	4.00
ford fairmont	0.00	2.00	2.00	4.00	0.00
ford mustang	0.00	5.00	0.00	7.00	1.00
ford pinto	0.00	0.00	2.00	1.00	0.00
honda accord	9.00	5.00	5.00	6.00	8.00
honda civic	8.00	4.00	3.00	6.00	7.00
lincoln continental	0.00	7.00	0.00	8.00	9.00
plymouth gran-fury	0.00	7.00	0.00	6.00	0.00
plymouth horizon	0.00	3.00	0.00	5.00	0.00
plymouth volare	0.00	4.00	0.00	5.00	0.00
pontiac firebird	1.00	0.00	0.00	7.00	8.00
volkswagen dasher	8.00	4.00	5.00	8.00	6.00
volkswagen rabbit	8.00	4.00	5.00	8.00	5.00
volvo dl	9.00	9.00	8.00	9.00	9.00

Note that ViSta cannot perform a Principal Components Analysis when there are more columns (subjects) than rows (automobiles), so we have choosen 14 of the subjects: About half the programmers and all of the non-programmers, including the company President, two people working in marketing and the grandmother of one of the programmers.

Principal Components Analysis

There are two important decisions that must be made when doing a principal components analysis: Should the analysis be based on correlations or covariances? And, how many compoenents are there?

Correlations or Covariances

Principal components may be computed using either correlations or covariances.

If the variables are all in the same units, as in the example used here, where everyone made judgements on the same scale, then you have the choice of either approach. If correlations are used then all variables are treated as equally important, whereas if covariances are used the importance of the variables is proportional to their variance.

If the variables are in different units, then one almost always has to base the analysis on correlations, since this involves standardizing all variables into the same, standard, units. Otherwise the variables are in incomparable units, and analysis of them doesn't make sense.

How many components:

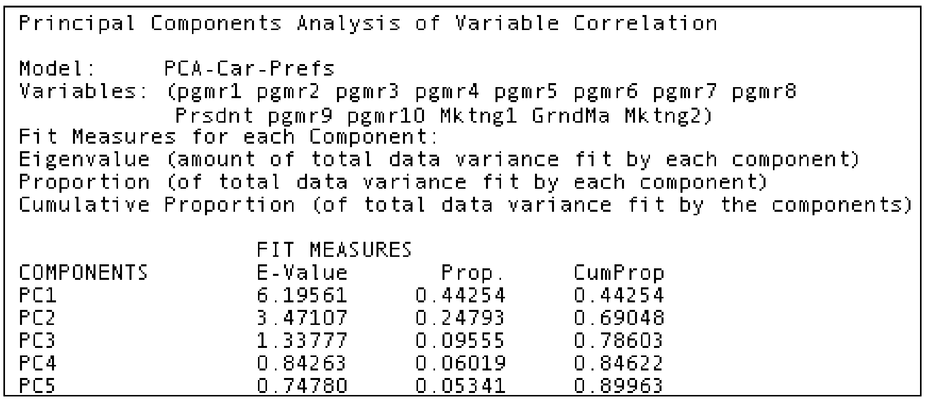
We now perform the principal components analysis and look at the report. The most important information is in the three Fit measures columns. The first two tell us the ammount and proportion of variance in the data that is fit by each principal component, and the third tells us the accumulated proportion fit by the several principal components up to and including the row the measure is in.

The complete set of principal components contains the same information as the raw data. Usually, you want to choose a smaller number of components for interpretation and subsequent use.

There are four commonly used criteria for deciding on the number of components:

1. It is the number of eigenvalues greater than 1.0 (when the analysis is based on correlation; use the average eigenvalue when the analysis is based on covariances).
2. Number of components required to account for a "meaningful" percentage of variance, usually 80-90%.
3. Plot the eigenv values and look for the "elbow".
4. See how many components are interpretable.

Fitted Variance: The first two criteria concern the amount of variance fit by the components model. Using these two criteria, we see that the first three components have eigenvalues greater than 1.0, and that three or four principal components account for a large fraction of the variance. We'll hold off on a decision until we look for an elbow and see about interpretation. But we probably don't need to keep more than four at the most.

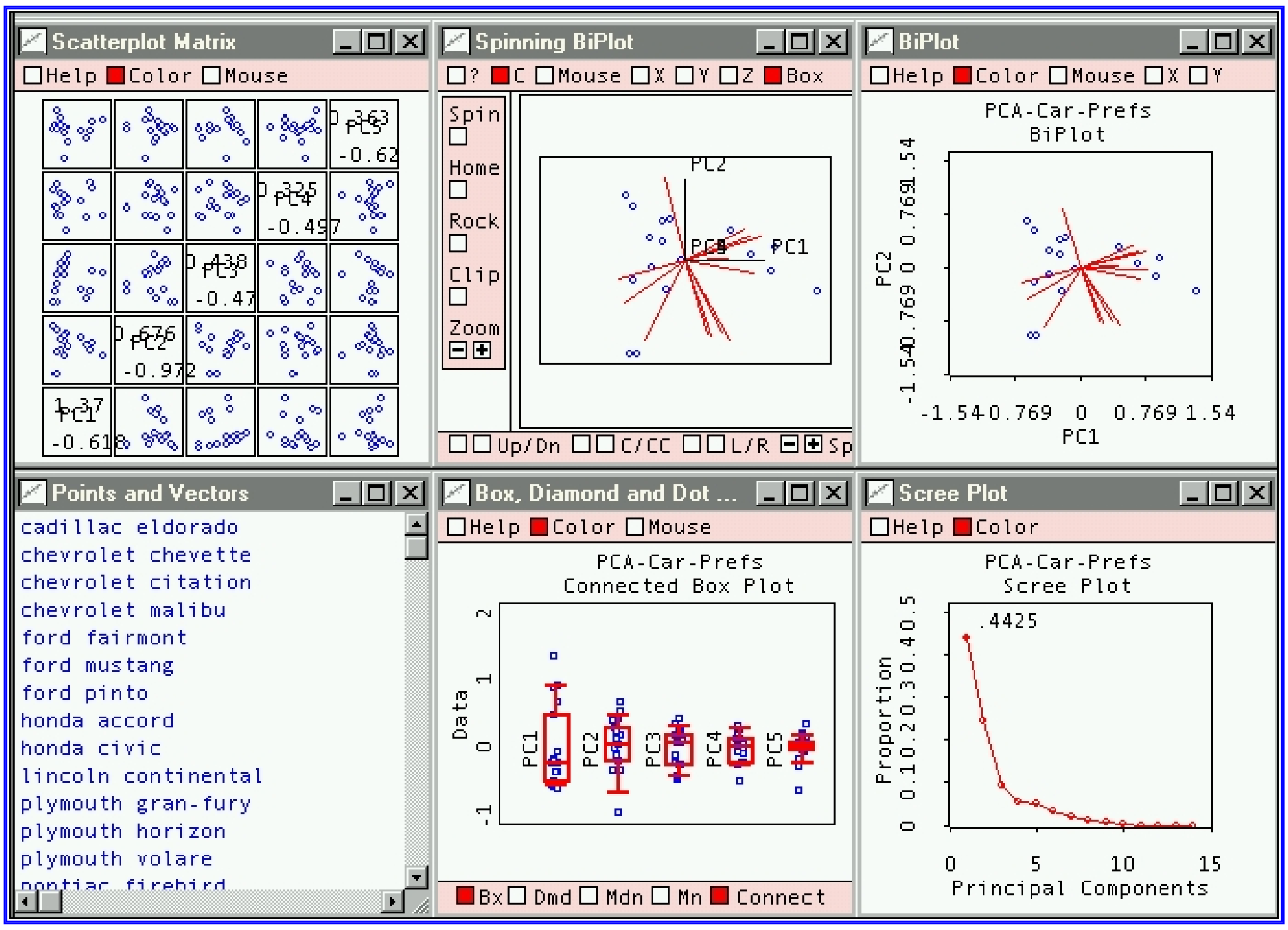


```
Principal Components Analysis of Variable Correlation

Model:      PCA-Car-Prefs
Variables: (pgmr1 pgmr2 pgmr3 pgmr4 pgmr5 pgmr6 pgmr7 pgmr8
            Prsdnt pgmr9 pgmr10 Mktng1 GrndMa Mktng2)
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                  6.19561      0.44254    0.44254
PC2                  3.47107      0.24793    0.69048
PC3                  1.33777      0.09555    0.78603
PC4                  0.84263      0.06019    0.84622
PC5                  0.74780      0.05341    0.89963
```

The spreadplot is presented below, it contains information relating to the third and fourth criteria.

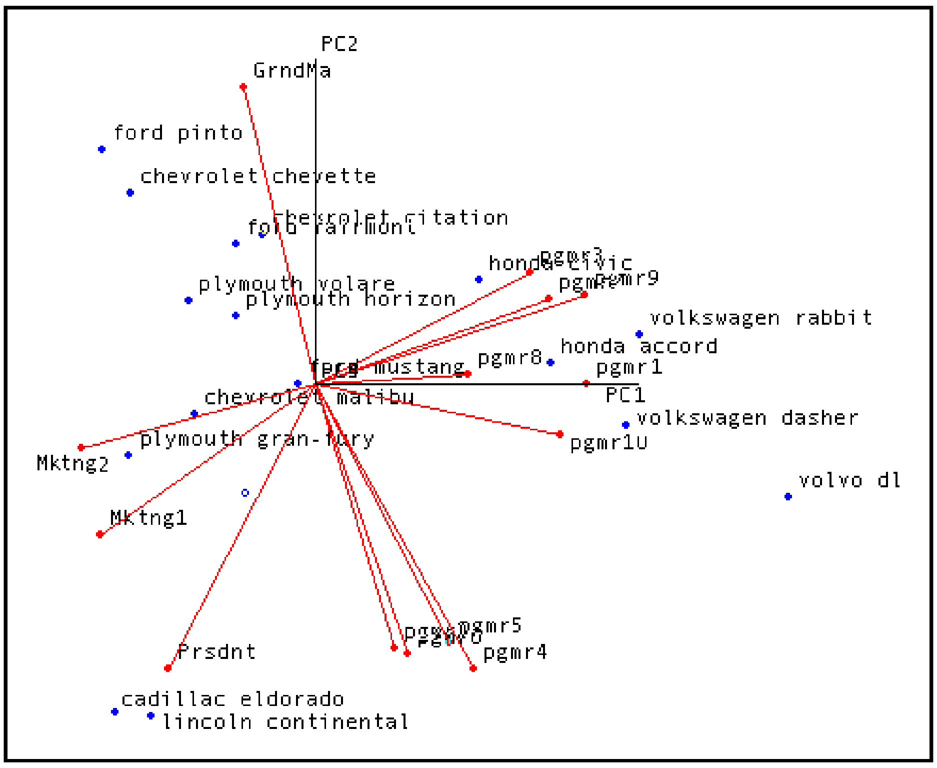


cadillac eldorado
chevrolet chevette
chevrolet citation
chevrolet malibu
ford fairmont
ford mustang
ford pinto
honda accord
honda civic
lincoln continental
plymouth gran-fury
plymouth horizon
plymouth volare
pontiac firebird

The most important plots here are the spin-biplot(upper-middle), the bi-plot (upper right) and the scree plot (lower right).

Elbow: The scree plot shows the porportion of variance fit by each component. We look at this plot to see if there is an "elbow" in the curve. Oftentimes we don't see any clear elbow. This one may have an elbow at three components. Having an "elbow" means that the increase in the amount of variance accounted for is relatively litle for each additional component. So this, and the preceeding information, leads us to conclude that not more than three may be needed to account for the "meaningful" variance in these data.

Interpretation: The two biplots show the same information except that one can be spun in three dimensions and the other is two-dimensional. This is the basic information that we interpret. For this reason the structure in the biplot is shown enlarged below.



For these data, where the vectors represent judges, and the points cars, a group of vectors pointing in the same direction correspond to a group of judges who have the same preference opinions about the automobiles. Thus, the judges whose vectors point towards 2 o'clock all have the same general likes and dislikes: What they like are imported cars and what they dislike are domestic cars. Note that these judges are all programmers. In contrast, the group of judges represented by the vectors pointing towards 5 o'clock (again, all programmers) like expensive cars, whether they are imported or domestic. Rather different is the one judge whose vector points towards 7 o'clock, the the two who point towards 8/9 o'clock. The first judge (who is the president of the company!) only likes expensive domestic cars, whereas the other two judges (who are in the marketing section of the company) like the "muscle cars". Finally, we have "GrandMa", who like the inexpensive cars!

If we look at the additional dimensions, there is very little more that we can understand, so we conclude that two components are what we need to interpret the data.