Principal Component Analysis

Principal Component Analysis (PCA) aims to identify patterns in data by reducing the dimensionality of multivariate data to a few key explanatory variables called principal components. PCA uses an orthogonal transformation to convert a data set of potentially correlated variables into a set of linearly uncorrelated variables called principal components. These vectors form an uncorrelated orthogonal basis set that retains as much information as possible. When there is a strong correlation between variables, the basis set can be used to reduce the dimensionality of the data by projecting it onto a smaller dimensional subspace.

The number of principal components is less than or equal to the number of original variables. The first principal component has the largest possible variance and each subsequent component has the largest possible variance given that it is orthogonal to the preceding components.

Example: Iris Data

The Iris data set contains measurements in centimeters for the variables sepal length, sepal width, petal length, and petal width, for 150 flowers from 3 species of iris, Iris setosa, versicolor, and virginica. The data was collected over several years by Edgar Anderson, who used the data to show that the measurements could be used to differentiate between different species of irises.

The following example will perform a principal component analysis on this data:

$with(Statistics)$:

Loading and visualizing the data

The Iris data set is available in Maple's datasets folder and can be imported using the ImportMatrix command:

```
IrisData := ImportMatrix(FileTools:-JoinPath(["datasets", "iris.csv"], base = datadir), skiplines
    = 1)
```

$$\begin{bmatrix} \textit{150 x 5 Matrix} \\ \textit{Data Type: anything} \\ \textit{Storage: rectangular} \\ \textit{Order: Fortran\_order} \end{bmatrix}$$

```
IrisLabels := ["Sepal length", "Sepal width", "Petal length", "Petal width"]:
```

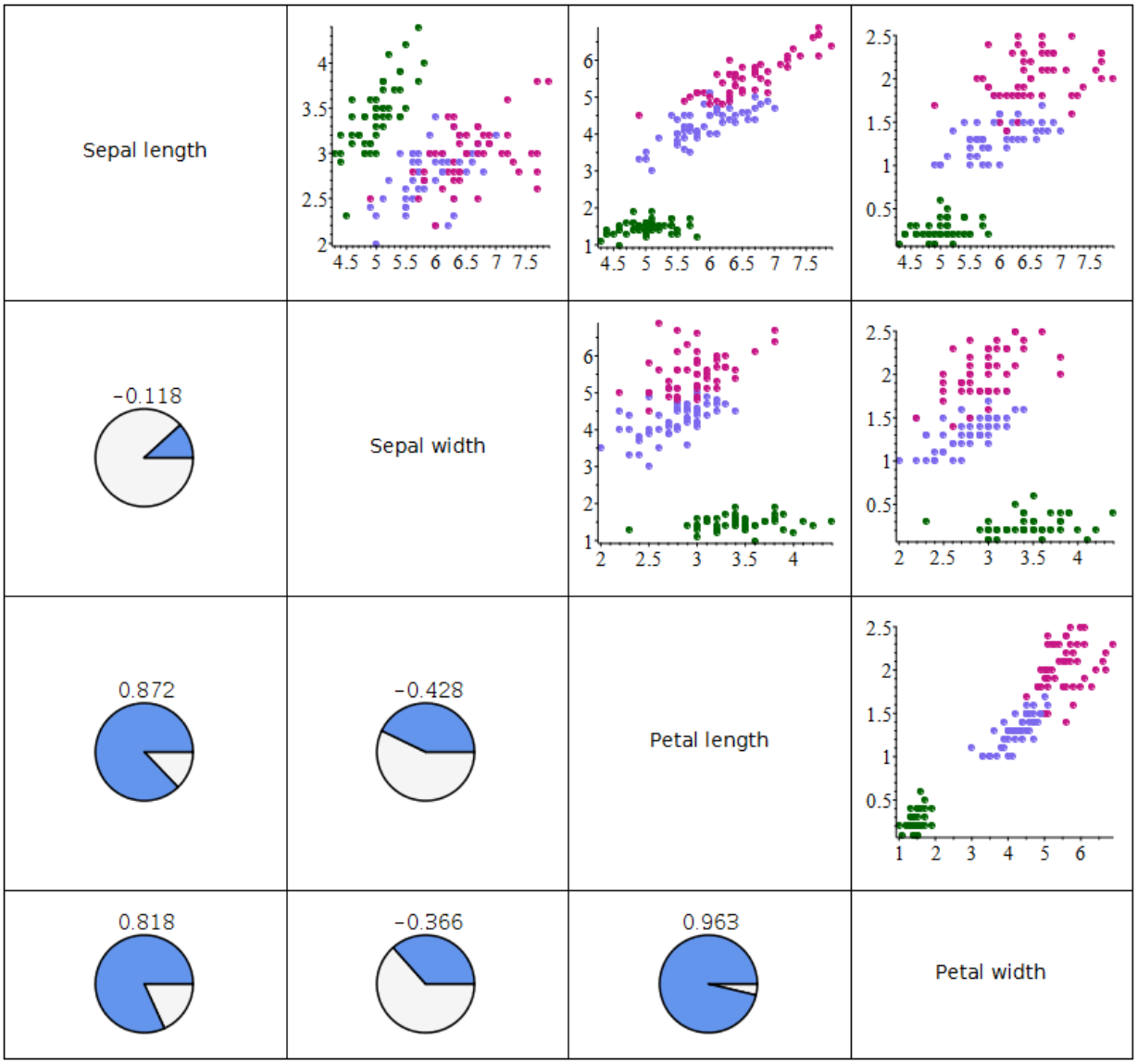
The DataSummary command returns more details on each column of data:

```
DataSummary(IrisData[.., 1..4], summarize = embed):
```

	1	2	3	4
mean	5.843333333333333	3.0573333333333315	3.758	1.1993333333333336
standarddeviation	0.8280661279778637	0.43586628493669793	1.7652982332594662	0.7622376689603467
skewness	0.3107121438818135	0.31471278918408596	-0.27121905735433127	-0.10159385768032718
kurtosis	2.4102558374011886	3.1597698087389894	1.5937676861769867	1.6528397108033237
minimum	4.3	2.0	1.0	0.1
maximum	7.9	4.4	6.9	2.5
cumulativeweight	150.0	150.0	150.0	150.0

In order to visually detect patterns between variables, the variables can be plotted against one another:

```
GridPlot(IrisData[.., 1..4], upper = [plots:-pointplot, colorscheme = ["valuesplit", IrisData[..,
    5], ["setosa" = "DarkGreen", "versicolor" = "MediumSlateBlue", "virginica"
    = "MediumVioletRed"]], symbol = solidcircle, symbolsize = 20], lower = '((x) → Statistics:-
    PieChart([" " = abs(x), " " = 1 - abs(x)], color = ["CornflowerBlue", "WhiteSmoke"],
    title = evalf[3](x), size = [100, 100]))', correlation = [false, true, false], labels
    = IrisLabels, width = 600, widthmode = pixels)
```



From the above, it can be seen that the "Petal Length" and "Petal Width" columns have a high level of correlation.

Principal Component Analysis

A principal component analysis can be run on the data to determine which variables explain the majority of the variability in the data.

$IrisPCA := PCA(IrisData[\,..,1..4],summarize)$:

```
summary:
Values   proportion of variance  St. Deviation
4.2282     0.9246                 2.0563
0.2427     0.0531                 0.4926
0.0782     0.0171                 0.2797
0.0238     0.0052                 0.1544
```

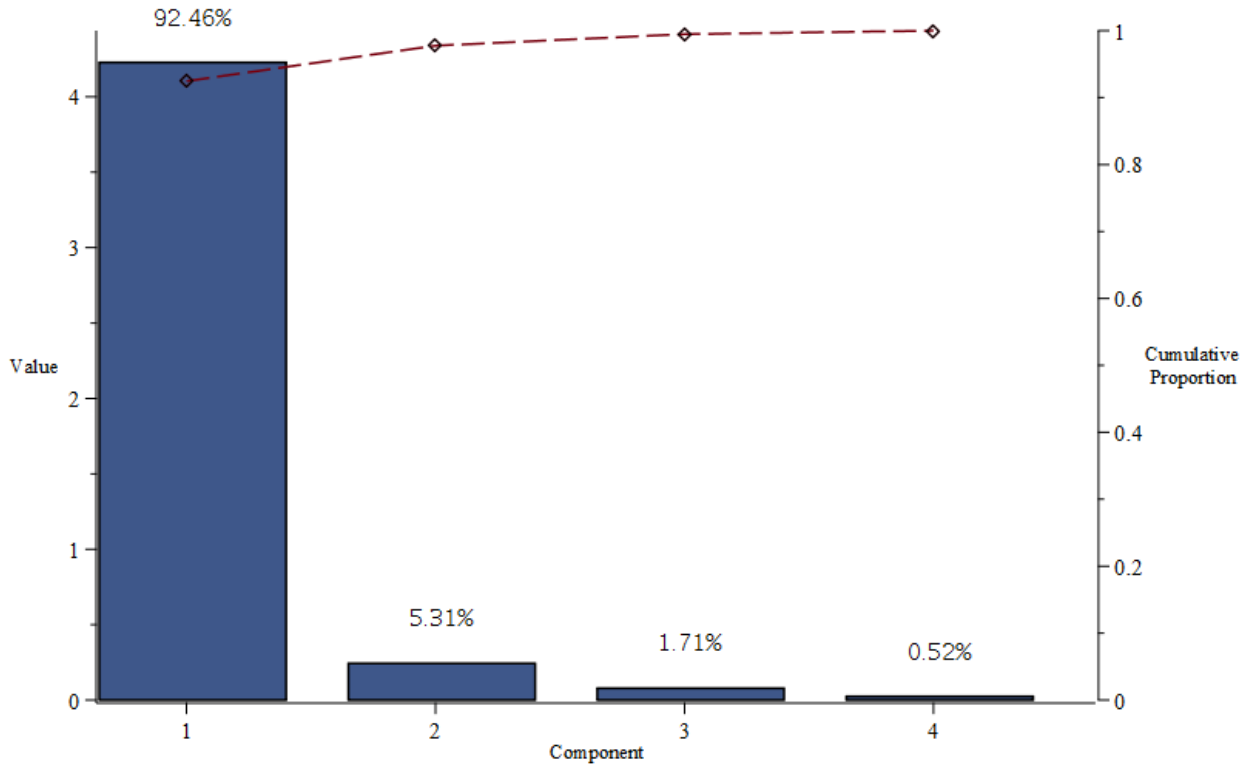
The principal component analysis command returns a record, which we can query in order to return the principal components, the rotation matrix, and details on the proportion of variance explained by each component. Note that this can also be seen by using the summarize option as above.

For example, the rotation matrix, or loadings for the components can be returned using the rotation option:

IrisPCA:-rotation

```
[[0.361386591785369, -0.656588771286843, 0.582029851306065, 0.315487192903975],
 [-0.0845225140645687, -0.730161434785026, -0.597910830100086,
 -0.319723103666129],
 [0.856670605949835, 0.173372662795857, -0.0762360758209631, -0.479838986994635
 ],
 [0.358289197151551, 0.0754810199174639, -0.545831432020076, 0.753657425264046]]
```

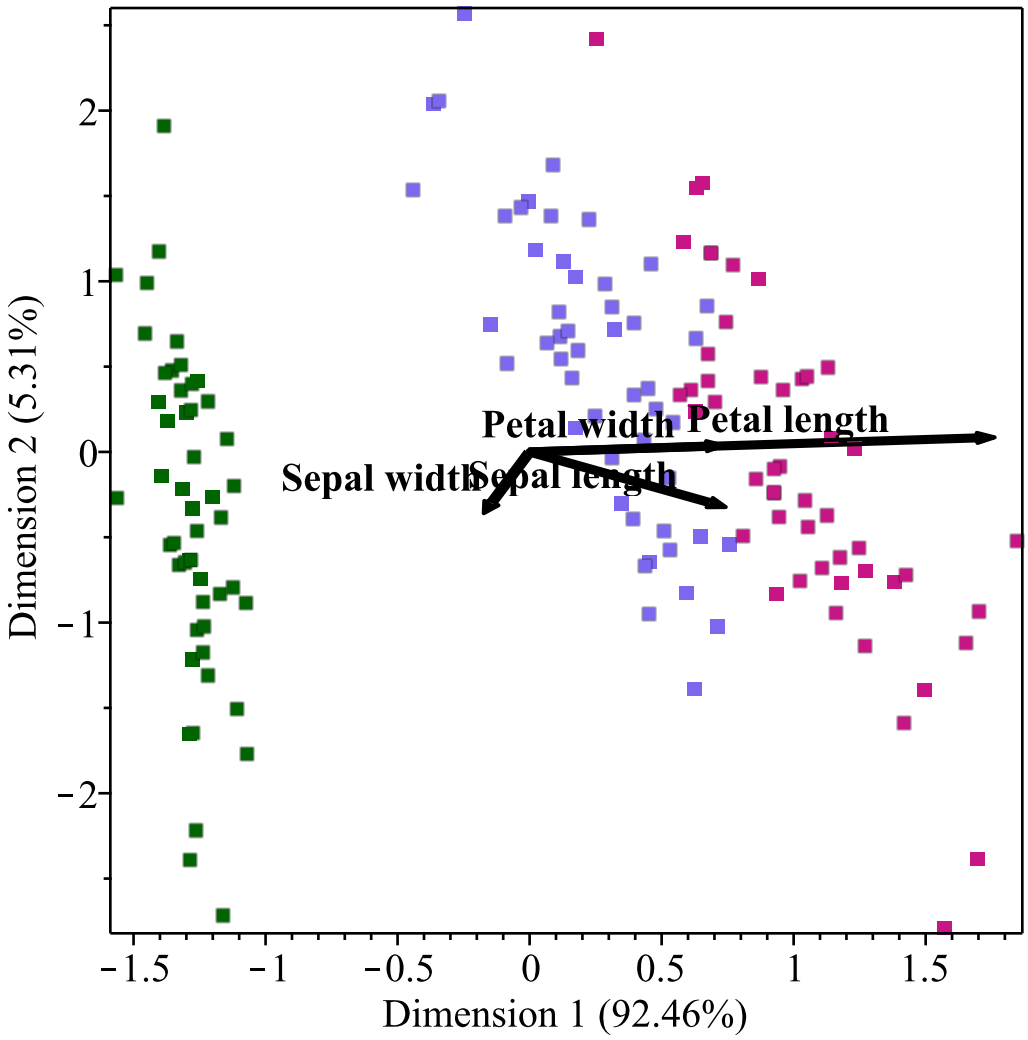
A ScreePlot is useful in visualizing the variance explained by each component:



From the scree plot, it can be seen that the first component accounts for 92.46% of the variance. The second component accounts for a much smaller fraction of the total variance, suggesting that only one component may be enough to summarize the data.

A Biplot can also be used to show the first two components and the observations on the same diagram. The first principal component is plotted on the x-axis and the second on the y-axis.

```
Biplot(IrisPCA, arrowlabels = IrisLabels, colorscheme = ["valuesplit", IrisData[ .., 5 ], ["setosa"
    = "DarkGreen", "versicolor" = "MediumSlateBlue", "virginica" = "MediumVioletRed"]])
```



From the biplot, it can be observed that petal width and length are highly correlated and their variability can be primarily attributed to the first component. Likewise, the first component also explains a large part of the Sepal length. The variability in Sepal width is more attributed to the second component.

References

Fisher, R.A. (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7, Part II, 179-188.