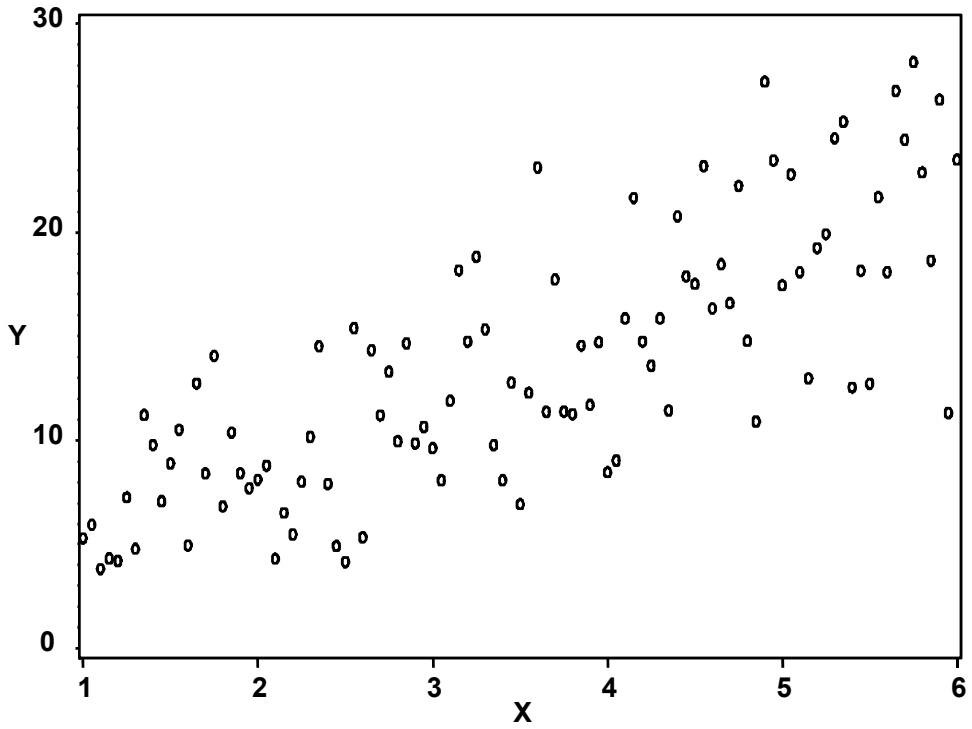
Introduction

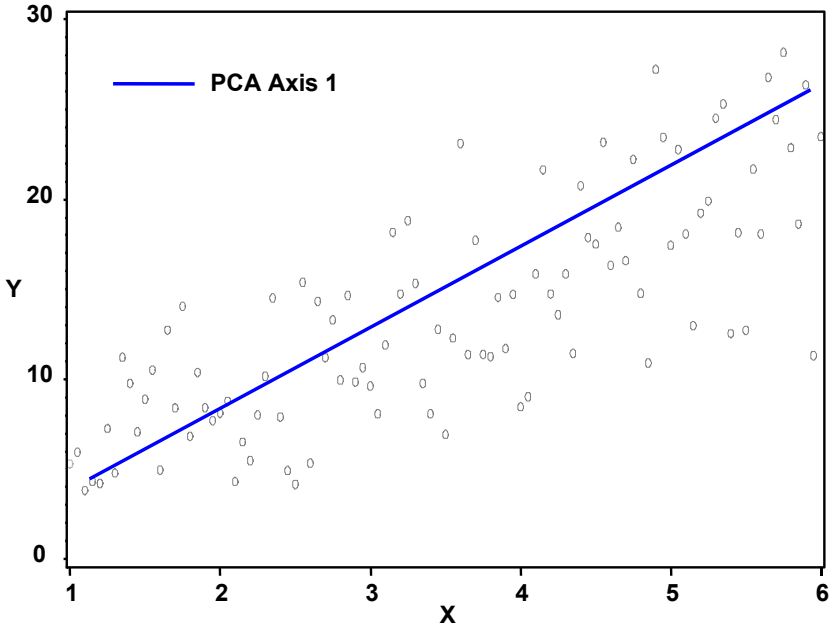
Unlike many "standard" statistical techniques, the goal of most multivariate procedures is not statistical inference, but rather to summarize the information contained in several variables. One method of achieving this goal is Principle Components Analysis or PCA. PCA is an important method to understand, because it is the underlying basis for many multivariate procedures. In PCA, the aim is to reduce a complex, multi-dimensional problem to a few understandable components. That is, to summarize several variables into a collection of a few which may have some practical meaning. A good way to understand this process is a graphical display of a simple case. In the example below, there are only two variables, X and Y. While a PCA would not typically be used with only two variables, it is easier to display the results with only two dimensions. Nevertheless, all the concepts described below may be extended to higher dimension systems.

A simple scatter plot of Y versus X shows that there is some correlation between the variables in the positive direction. Such correlation is common in real multivariate data and PCA is often used to help mitigate the effects of such collinearity.

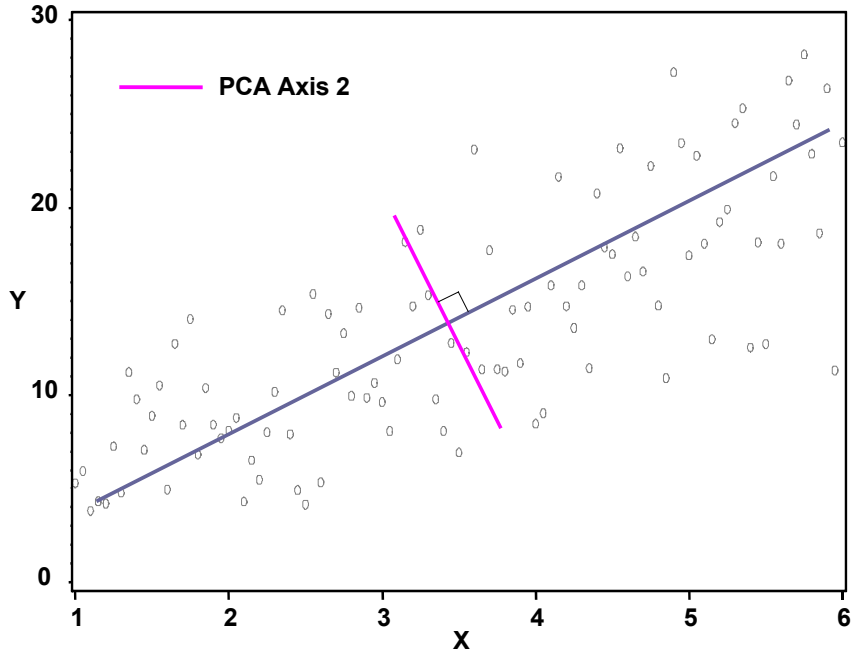


PCA works by defining linear combinations of the variables (axes) such that they describe the largest amount of variability in the "data cloud". In this case, most of the variability lies in the Y variable (ranging from 0 to 30). We might expect a linear combination of X and Y which captures most of the variability to be a line that runs diagonally upward from left to right. The figure below shows this axis which is labeled PCA axis 1 or principle component one. For these data, the axis may appear similar to a linear regression line. However, regression seeks to minimize the distance between data

points and the fitted line, e.g. through the method of least squares, while the PCA seeks to define the line (linear combination of variables) which runs through the longest dimension of the data, i.e. the line that covers the most variability.



The next step in the PCA is to define a second component or PCA axis 2. This axis must satisfy the conditions that it cover the next longest dimension of the data and that it be perpendicular or at a right angle with the first axis. This perpendicularity or orthoganality is important because it ensures that the two PCA axes are independent from each other. In general, all PCA axes will be orthogonal (independent) to each other. The second axis for this example is shown below.



The number of PCA axes that can be defined in a given data set will be equal to the number of variables it contains. In our example, the data had two variables, so the PCA is

limited to two axes. Within each axis, we are interested in two pieces of information: 1) the variability accounted for by the axis, and 2) what linear combination of variables make up the axis. These are referred to as the eigenvalues and eigenvectors (or loadings), respectively. The eigenvalues for this example are:

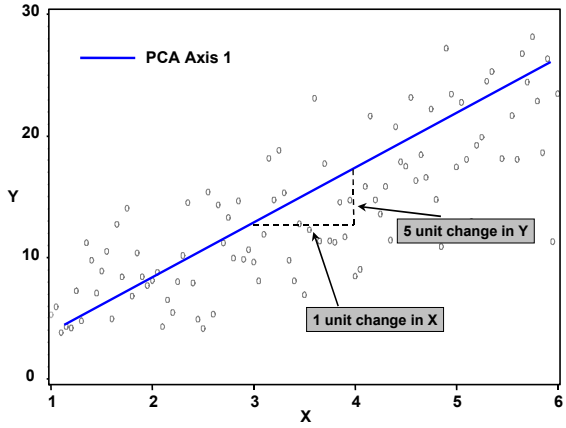
Axis	Eigenvalue	Proportion	Cumulative
1	40.4234898	0.9799	0.9799
2	0.8294980	0.0201	1.0000

Here, the first axis has an eigenvalue or variance of 40.4 which accounts for 97.99 % of the total variability in the data. The second axis is much less important with a variance of 0.83 and 2% of the variability. Note that the first axis will always account for the most variability in the data, and the proportion of variance accounted for by each subsequent axis will sequentially decrease. In a data set with many PCA axes, the first two or three axes should typically account for most of the variability. This indicates that the information available in the data can be accurately summarized into two or three linear combinations of the original variables. If the first few axes cannot account for most of the variability, then PCA is unlikely to provide any help for summarizing the data.

The second item of interest in PCA is what combination of the variables are used to derive the important axes. This information is contained in the eigenvector of each axis. The eigenvectors for this example are:

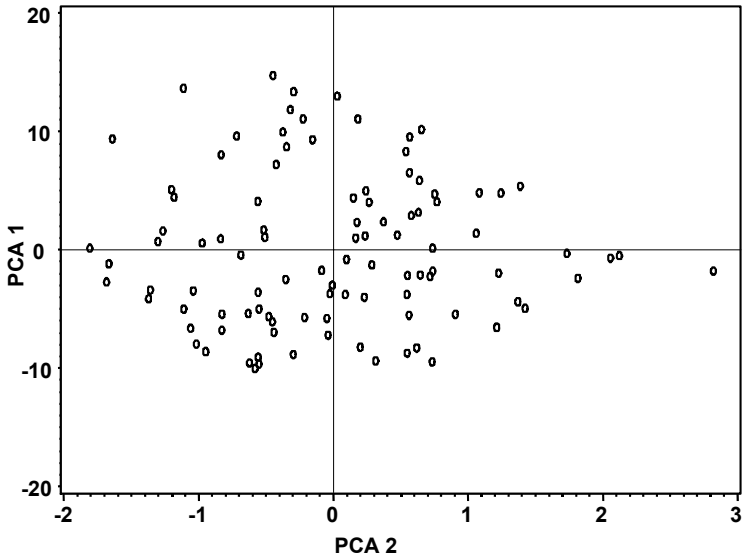
	PCA 1	PCA 2
Y	0.983231	-.182363
X	0.182363	0.983231

In the PCA columns, the coefficients or loadings for the Y and X variables are provided. As noted earlier, axis 1 accounts for most of the data variability. This axis is defined as 0.98*Y + 0.18*X. Axis 1 is said to be dominated by or representative of variable Y. In fact, the ratio of Y to X is about 5 to 1. This can be seen graphically by looking at axis 1 again.



The axis defines about 5 units change in Y for every unit change in X. From these results we could conclude that 1) PCA 1 accounts for most of the data variability, and 2) the axis is

mostly composed of variable Y. Thus, the variability in these data is mainly due to the variable Y. Also note that these linear combinations of X and Y are now independent and uncorrelated. This can be observed by computing the linear combinations for each observation (PCA scores) and plotting the resulting PCA axes against each other (referred to as a biplot):



These plots are sometimes useful for identifying grouping or structure within a data set. In this case, however, no such structure is evident.