Principal Components

There are a variety of ways to reduce number of variables used graphics and models. The goal is to transform the variables into fewer new variables that retain most of the variation (information) in the original variables. The principal components approach computes new variables as linear combinations of the original variables. This method produces new variables that are linearly independent (so have zero correlation) and are listed in descending order in terms of their variance.

Researchers have two basic choices to make when using principal components. The first step in constructing principal components is to center the variables (subtract each variable's mean so the new mean is zero). The first choice is whether or not to scale the variables. This divides the centered variables by their standard deviations. (We can easily center and scale a matrix of continuous variables with R's scale function.) The choice is usually, yes, scale the variables, when the variables have different units of measures. Then after scaling we can interpret the coefficients of the linear combinations and the values of the new variables in terms of unitless standard deviation units.

When the original variables all have same units of measures, choosing no means that the new variables will be in same units of measure as the original data. This may be preferred for interpretation convenience. The choice of scaling or not is equivalent to using either the correlation matrix or the covariance matrix. respectively, as the basis for constructing the principal components.

The second choice is selecting the principal component to use. Typically researchers choose the the principal components with largest variances. (There some cases with the key information ends up in low variance principal components.) Most of the time the choices is about how many principal components to use for graphics or models. In R a screeplot shows the variances for the principal component. Fairly often there is point of diminishing returns for including more variables. That is, adding another variable increases the total variation by very little. In the absence of additional consideration, this seems reasonable heuristic for making a choice.

Regression can handle a large number of principal components. While with graphics it is common to see graphics showing just the first two principal component, it is possible to view several variables, for example, using scatterplot matrices or parallel coordinates. In scientific studies the first two or three principal components often end up representing known dominant source of variation and the additional principle components provide the opportunity to see less dominant structure that may not have been previously identified.

The coefficients for the linear combinations are often call loadings.