Principal Components: Not Just Another Factor Analysis

Statnews #49

Cornell Statistical Consulting Unit

Created February 2002. Last updated October 2020

Introduction

Principal components analysis is a technique for reducing the dimension of a multivariate dataset while retaining as much of the variance in that dataset as possible. The resulting transformed variables, or "principal components", can be used to visualize high-dimensional observations in two dimensions, as predictor variables in a regression model, or to define an index which combines information from all of the variables.

In a dataset with n observational units and d variables measured on each unit, up to d principal components can be computed (assuming there are more observations than variables, i.e. $n > d$). By construction, the principal components are uncorrelated, and typically just a few principal components are retained for subsequent analysis or data visualization.

Mathematical description and example

For simplicity, suppose there are four variables measured on each observational unit, x_1, x_2, x_3, x_4. A principal component is defined by

$$z = w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4.$$

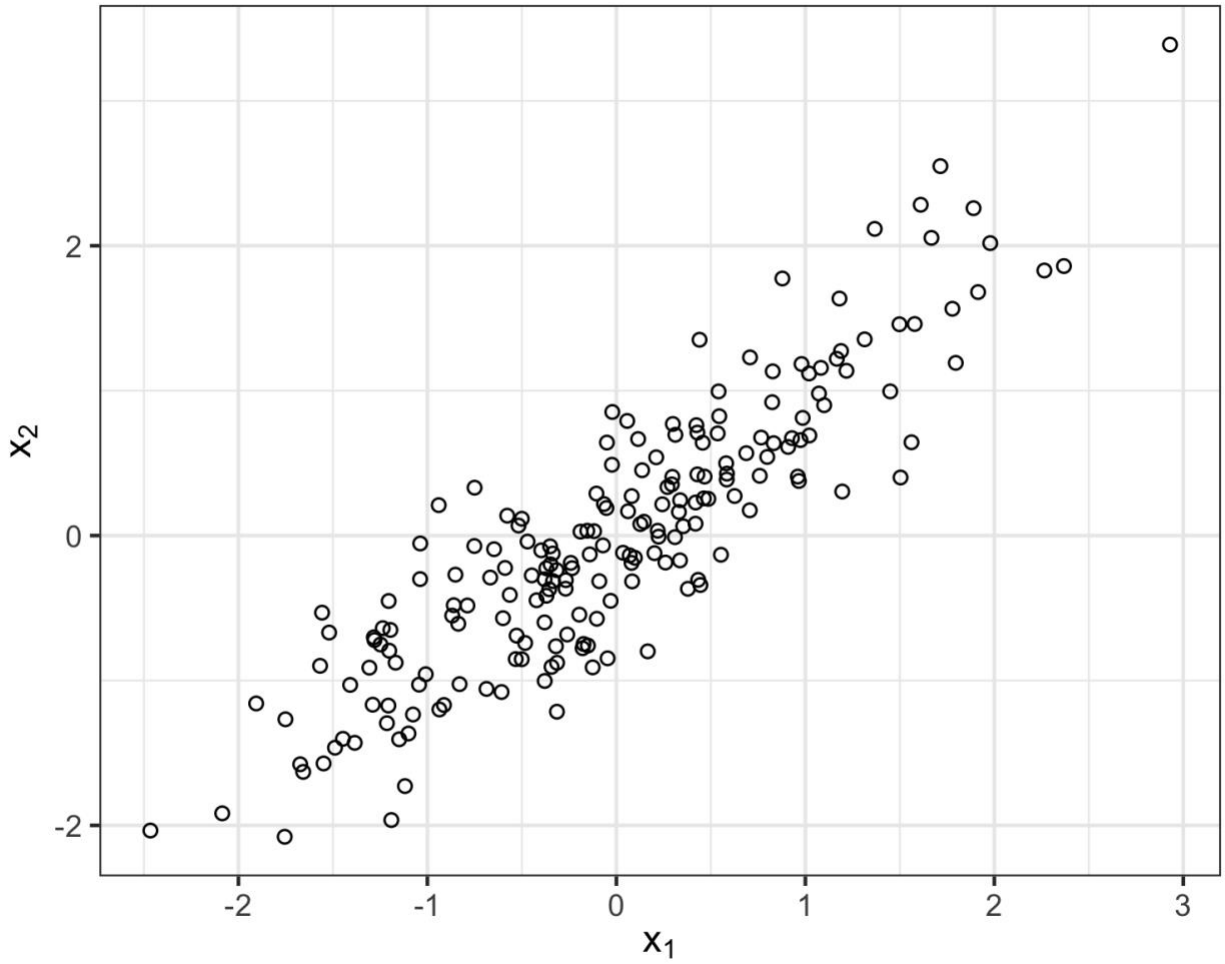
The values w_1, w_2, w_3, w_4 are weights called *loadings*. The loadings are chosen to maximize the sample variance of the resulting z values. Thus, the principal component is a weighted combination of the original variables that contains as much variance as possible from those original variables.

With 4 variables (and at least 5 observational units), we can compute 4 principal components. The first principal component is defined by the choice of weights with the maximum possible sample variance. Subsequent principal components are found by computing new weights so that all the principal components are uncorrelated and have the maximum possible sample variance after accounting for the variance in the previous principal components.

With d original variables, the sum of the variances of all d principal components will be equal to the sum of the variances of the original variables. But since each principal component is a

weighted sum with the maximum possible variance, it is often the case that a large proportion of the total variance will be captured by the first few principal components.

For illustration, consider the following example with just $d = 2$ variables (in practice principal components analysis would likely not be beneficial with only two variables). The $n = 200$ simulated observations of these two variables are displayed in Figure 1.



The sample variances of x_1 and x_2 are 0.91 and 0.88, respectively, for a total variance of 1.79. The first principal component has weights $w_1 = 0.71, w_2 = 0.7$ and sample variance 1.68, which is about 94 percent of the total variance in the original two variables. Figure 2 plots the principal components for these same $n = 200$ observations.

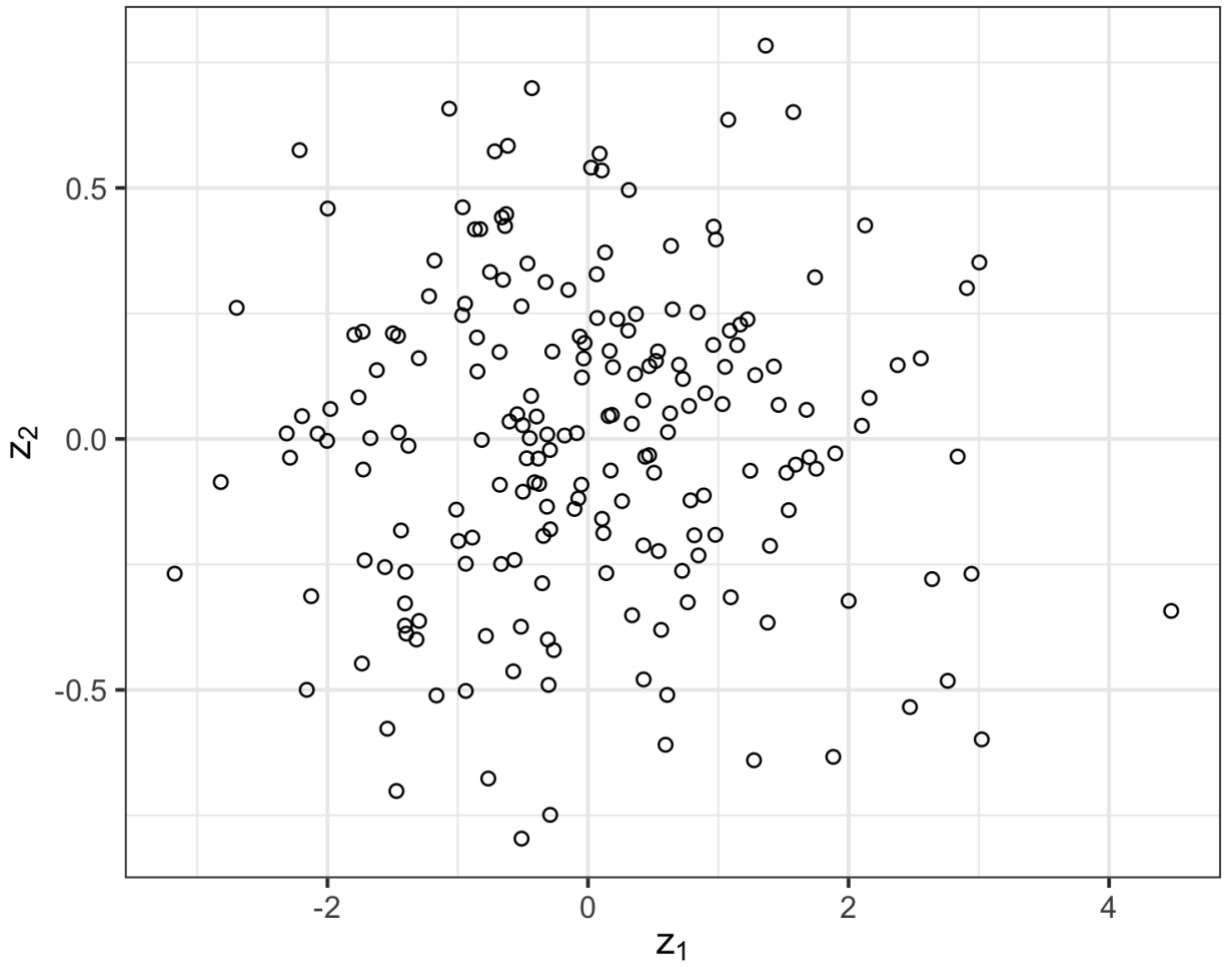


Figure 2: Principal component scores z_1, z_2 for the simulated observations.

Principal components vs. factor analysis

Principal components analysis and factor analysis are similar to each other in two ways. First, both are data reduction techniques that create a new smaller and manageable set of variables (principal components for one and factors for the other). Second, one of the extraction methods (i.e., principal axis method) used for factors in factor analysis is actually the same as the one used for principal components.

But there are also important differences between principal components and factor analysis. Conceptually, factor analysis requires a meaningful model that assumes that the correlation among variables is due to the fact that they are a manifestation of one or more common underlying factors (see StatNews #48: What is factor analysis?). In other words, factor analysis supposes that the observed variables are a result of a weighted combination of unobserved, latent factors, and attempts to estimate those weights and latent factors. In contrast, principal components analysis finds weighted combinations of the *observed variables* with maximum sample variance.

Furthermore, because factor analysis is a model similar to regression, we expect that, for each observed variable, some of the variability will be explained by the model and some will not. In contrast, in principal components analysis, all variability in the original variables will be explained by the components.