FACTOR ANALYSIS FOR NON-NORMAL VARIABLES

Ab Mooijaart

Leiden University

Factor analysis for nonnormally distributed variables is discussed in this paper. The main difference between our approach and more traditional approaches is that not only second order cross-products (like covariances) are utilized, but also higher order cross-products. It turns out that under some conditions the parameters (factor loadings) can be uniquely determined. Two estimation procedures will be discussed. One method gives Best Generalized Least Squares (BGLS) estimates, but is computationally very heavy, in particular for large data sets. The other method is a least squares method which is computationally less heavy. In one example the two methods will be compared by using the bootstrap method. In another example real life data are analyzed.

Key words: factor analysis, nonnormality, third order cross-products, unique factor loadings, BGLS estimates, bootstrap.

1. Introduction

In factor analysis only second-order cross-products (like covariances or correlations) are analyzed. Although no explicit reason for this restriction is given, the implicit argument seems to be: Whenever the variables are normally distributed, the covariance matrix is a sufficient statistic for the multivariate normal distribution and so higher order cross-products do not add any information (Bentler, 1983b). However, parameters (factor loadings and unique variances) cannot be identified uniquely from covariances only. For a detailed discussion of the relationship between identifiability and the normality assumption in factor analysis the reader is referred to Reiersøl (1950), Rao (1966) and Kagan, Linnik and Rao (1973). Very roughly (i.e., by omitting some conditions), their result is: If factor loadings are not identifiable then some variables, factors or residuals must be normally distributed and the reverse; that is, if factor loadings are identifiable then some variables are not normally distributed. A conclusion is that if variables are normally distributed then parameters cannot be identified in general. However, if the variables are not normally distributed then the parameters may be identified.

Up to now there are two ways of still getting unique parameters by analyzing the covariance matrix only. The oldest way is by rotating a solution to some structure with specific features. A more recent approach is by fixing some parameters and finding unique estimates for the remaining parameters. (This approach is very common in LISREL, as developed by Jöreskog & Sörbom, 1981). The rotation solutions (e.g., simple structure, or Varimax) are mainly concerned with methods which facilitate interpretation of factors. These methods merely decompose the covariance matrix in a specific way and do not specify some specific structural equation. By fixing some factor loadings factor analysis is in fact a structural model. A problem with this approach is that a theory is necessary for
specifying which parameters have to be fixed and fixed to what value (mostly zero or one).

The works of Reiersol (1950), Rao (1966) and Kagan, Linnik and Rao (1973) give a hint in which direction we have to search for finding unique estimates: dropping the normality assumption. However, the covariance matrix is then no longer a sufficient statistic, which means that higher order cross-products give additional information about the underlying structure. In this paper second and third order cross-products will be analyzed and conditions will be given under which circumstances parameters are identified.

The idea of normally distributed variables in factor analysis was first proposed by Lawley (1940), who developed maximum likelihood estimates based upon the assumption that the variables are normally distributed and that the covariance matrix follows the Wishart distribution. After computational difficulties were solved (mainly by Jöreskog, 1967), the popularity of statistical estimation in factor analysis with normally distributed variables has increased sharply. The main reason for this is that by ML factor analysis one can test the model and find standard errors for the estimates.

However, more recently, it has been questioned whether the assumption of normality is realistic or not. This is an important point because it is known that the maximum likelihood method may be highly non-robust against violation of the normality assumption (Layard, 1972, 1974). So if the normality assumption is violated then the likelihood ratio test gives incorrect results. (For more discussions about this point see Browne, 1982, Steiger & Hakstian, 1982, Boomsma, 1983, Bentler, 1983a, and de Leeuw, 1983).

Jöreskog and Goldberger (1972) proposed as an alternative to the maximum likelihood method a generalized least squares (GLS) procedure which has under the assumption of normality the same asymptotic properties as the maximum likelihood procedure. Browne (1974) generalized their results to nonnormal variables. He showed that if the weight matrix in the GLS function is a consistent estimator of the population covariance matrix and the variables have zero fourth order cumulants (that holds for normal distributed variables, but may also hold for other types of distributions), estimators have the same asymptotic properties as in the ML procedure. Although this result is rather theoretical, it shows that not only ML estimates are BAN estimates. Other examples of estimates in factor analysis with the same asymptotic properties as the maximum likelihood estimates are given by Swain (1975).

Browne (1982, 1984) generalized his earlier results to elliptical distributions. Elliptical distributions may have nonzero fourth order cumulants but all marginal distributions have the same kurtosis. The class of elliptical distributions is more broad than the class of normal distributions; still, the question remains whether this assumption is realistic.

Unfortunately there are few psychological theories about how the variables are distributed. It seems reasonable (and also for identification of the parameters, as discussed above) to apply models in which weak distributional assumptions are made. Therefore one estimation procedure we shall use is an asymptotic distribution free (ADF) estimation method, by which it is still possible to test the model and to find standard errors under some very mild asymptotic assumptions.

As said before, we will utilize expectations of third order cross-products, besides expectations of second order cross-products. However, the ADF method using third order cross-products may become impracticable if the number of variables is too large. Firstly, the computational burden may be too large for many variables, i.e., the order of the covariance matrix of the cross-products increases sharply with the number of variables. Secondly, the estimation of the covariance matrix mentioned above may be very unstable, in particular for small sample sizes. (This last point is being investigated at the moment.) In these cases other estimation procedures have to be applied. In this paper we show by an example how the least squares method works out and what the results are of applying a bootstrap procedure.