Robust Principal Component Analysis by Projection Pursuit

Heinrich Fritz and Peter Filzmoser
Department of Statistics and Probability Theory
Vienna University of Technology

Abstract: Different algorithms for principal component analysis (PCA) based on the idea of projection pursuit are proposed. We show how the algorithms are constructed, and compare the new algorithms with standard algorithms. With the R implementation *pcaPP* we demonstrate the usefulness at real data examples. Finally, it will be outlined how the algorithms can be used for robustifying other multivariate methods.

Keywords: Projection pursuit, Robustness, Principal component analysis, Multivariate methods, pcaPP

1 Introduction

Many multivariate statistical methods are based on a decomposition of covariance matrices. For high-dimensional data this approach can be computationally intensive, especially if the involved covariance matrices should be estimated in a robust way. Moreover, if the sample size is lower than the dimension, additional problems with robust covariance estimation will arise.

An alternative approach for obtaining robust multivariate methods is projection pursuit (Huber, 1985). For example, in PCA the first component is defined as that direction maximizing a measure of spread of the projected data on this direction. If a robust spread measure is considered, the resulting PC is robust. Thus, robust estimation is done only in one dimension, namely in the direction of the projected data.

A non-trivial task is finding the direction which maximizes an objective function, like a robust spread measure for robust PCA. In this context, Croux and Ruiz-Gazen (2005) suggested to use each observation for the construction of candidate directions. We will extend this idea and introduce other algorithms. In a straightforward manner we can also obtain other (robust) multivariate methods.

2 Extensions of the Algorithm of Croux and Ruiz-Gazen (2005)

Croux and Ruiz-Gazen (2005) suggest to use as candidate directions for the first PC all directions from each data point through the center of the data cloud, estimated e.g. by the L_1-median. Subsequent PCs are estimated in a similar way, but the search is done in the orthogonal complement of the previously identified PCs. However, due to its construction, this algorithm may not be very precise for data sets with low sample size n or where n/p is low, with p being the number of variables. And there is yet another problem: By construction, the direction is determined by one of the data points. When the data are projected to the orthogonal complement, the projection of this data point is zero. This can lead to implosion of the scale estimator if p is sufficiently high.

To avoid these drawbacks one can add an updating step which is based on the algorithm for finding the eigenvectors. The drawbacks of the algorithm of Croux and Ruiz-Gazen (2005) can also be avoided by taking, in addition to the n data points, other candidate directions for maximizing the objective function. These directions are randomly generated: Generate n^+ data points with p-dimensional multivariate standard normal distribution, and project the data to the unit sphere. The directions of each generated data point through the origin are the new random directions, and by definition they have norm one.

3 Grid Algorithm

The optimization is always done in a plane rather than in the p-dimensional space. The first step is to sort the variables in descending order according to the largest scale. Then the optimization is done in the plane spanned by the first two sorted variables, where the candidate directions are constructed by dividing the unit circle into a regular grid of segments. A second approximation of the projection direction is then found by maximizing in the plane formed by the first and the third sorted variable. This process is repeated until the last variable has entered the optimization, which completes the first cycle of the algorithm. In a second cycle each variable is in turn again considered for improving the maximal value of the objective function. The algorithm terminates after a fixed number of cycles or when the improvement is considered to be marginal.

4 Robust Multivariate Methods

Above we described algorithms for estimating the (robust) PCs. We can use these for building a (robust) covariance matrix, which then can be plugged in into multivariate methods like factor analysis, canonical correlation analysis or discriminant analysis. On the other hand, some of the multivariate methods can be reformulated as a projection pursuit method, and the above algorithms could be applied. This approach was used for robust continuum regression (Filzmoser et al., 2006).

Croux, C., and Ruiz-Gazen, A. (2005). High breakdown estimators for principal components: the projection-pursuit approach revisited. *Journal of Multivariate Analysis*, 95, 206-226.

Filzmoser, P., Serneels, S., Croux, C., and Van Espen, P.J. (2006). Robust multivariate methods: The projection pursuit approach. In: Spiliopoulou, M., Kruse, R., Nürnberger, A., Borgelt, C., and Gaul, W. (Eds.), *From Data and Information Analysis to Knowledge Engineering*, Springer-Verlag, Heidelberg-Berlin. To appear.

Huber, P.J. (1985). Projection pursuit. *The Annals of Statistics*, 13 (2), 435-475.