Diagnostics for Robust Multivariate Analysis

Stefan Van Aelst[1] and Greet Pison[2]

[1] Ghent University,
Department of Applied Mathematics and Computer Science,
Krijgslaan 281 S9,
B-9000 Gent, Belgium
[2] University of Antwerp
Department of Mathematics and Computer Science,
Universiteitsplein 1,
B-2610 Wilrijk,
Belgium

Keywords: Empirical influence function, Diagnostic plots, Robust distances.

1 Abstract

Principal component analysis, canonical correlation analysis and factor analysis are popular methods for analyzing multivariate data. Recently, robust versions of these methods have been proposed and investigated (Croux and Haesbroeck 2000, Croux and Dehon 2001, Pison et al. 2002). Influence functions and corresponding empirical influence functions for these methods have been derived. However, there does not yet exist a graphical tool to display the results obtained from the robust data analysis in a comprehensive way. Therefore we now construct such a diagnostic tool based on empirical influence functions.

2 Diagnostics

Data points that are outlying are not necessarily influential observations for the multivariate model. This is very similar to regression analysis where a huge outlier in the cariers can have a very small standardized residual and therefore is called a *good leverage point* (Rousseeuw and van Zomeren 1990). It is not recommendable to downweigh or delete such points because these observations improve the accuracy of the estimates. In regression analysis a diagnostic tool proposed by (Rousseeuw and van Zomeren 1990) allows to identify very quickly the outliers, the good leverage points and the bad leverage points. Similarly, we now also construct such a graphical tool for robust multivariate methods. Let us first look at principal component analysis.

To construct the diagnostic plots we will use the empirical influence of each observation on the eigenvalues and eigenvectors. To obtain these empirical influences we substitute robust estimates for the unknown parameters in the influence functions of the functionals corresponding to the classical estimators.

This kind of empirical influence function was already proposed in Pison et al. (2002) who show that it indeed detects the most influential data points. To estimate the unknown parameters we use the MCD estimator (Rousseeuw 1984) or S-estimators (Davies 1987, Rousseeuw and Leroy 1987) which are highly robust estimators that are easy to compute.

For each observation an overall influence for the eigenvalues and eigenvectors is computed by taking the euclidean norm of the influences of the components. To detect influential points we also need a cutoff value. Therefore we generate 100 datasets from a multivariate normal distribution which has as correlation matrix the robust correlation matrix estimate of the original data. For each of these datasets we compute the empirical influences and determine the cutoff value as the 95% percent quantile of these empirical influences.
Finally, we plot the overall empirical influence of the observations together with the corresponding cutoff value versus the robust distances of the observations with their corresponding cutoff value. This diagnostic plot then divides the observations into regular points, non-outlying influential points, influential outliers and non-influential outliers. In a second step the influential outliers can be downweighted in a reweighted classical principal component analysis which will then yield more efficient estimates of the parameters.
Similar plots will be constructed for canonical correlation analysis and factor analysis.

References

C. Croux and G. Haesbroeck (2000). Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika*, 87, 603-618.

C. Croux and C. Dehon (2001). Analyse canonique basee sur des estimateirs robustes de la matrice de covariance. *La Revue de Statistique Appliquée*, to appear.

L. Davies (1987). Asymptotic behavior of S-estimators of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, 15, 1269-1292.

R.A. Johnson and D.W. Wichern (1998). *Applied Multivariate Statistical Analysis.* Fourth Edition, Prentice Hall, New Jersey.

G. Pison, P.J. Rousseeuw, P. Filzmoser, and C. Croux (2002). Robust factor analysis. *Journal of Multivariate Analysis*, to appear.

P.J. Rousseeuw (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79, 871–880.

P.J. Rousseeuw and A.M. Leroy (1987). *Robust Regression and Outlier Detection.* Wiley-Interscience, New york.

P.J. Rousseeuw and B.C. van Zomeren (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85, 633-651.

Please fill in this form and mail it together with your abstract.

My abstract fits best to topic number 4

(**"PLEASE INDICATE HERE THE NUMBER FROM THE LIST OF TOPICS BELOW
WHICH BEST FITS TO YOUR ABSTRACT"**.).

List of Topics:

1. Algorithms
2. Applications
3. Biostatistics
4. Computing and graphics
5. Data analysis
6. Data mining
7. Economics, finance
8. Efficiency and robustness
9. Functionals and bias
10. Fuzzy statistics
11. Geostatistics
12. Inference for robust methods, model testing
13. Location depth and regression depth
14. Multivariate methods
15. Neural networks
16. Rank-based methods
17. Regression quantiles, trimming
18. Robust covariance
19. Robust designs
20. Robust regression
21. Time series analysis
22. Wavelets
23. Other (please specify)