Robust Location and Scatter Estimators for Multivariate Analysis

Valentin Todorov
Austro Control GmbH
valentin.todorov@chello.at

Abstract: This talk reviews the most popular robust alternatives of the classical multivariate location and scatter estimates and discusses their application in the multivariate data analysis with main emphasis on the availability in R.

Keywords: Robust Estimation, Multivariate Analysis, Minimum Covariance Determinant Estimator, MCD

1 Introduction

The estimates of the multivariate location vector $\boldsymbol{\mu}$ and the scatter matrix $\boldsymbol{\Sigma}$ are a cornerstone in the analysis of multidimensional data, since they form the input to many classical multivariate methods. The most common estimators of the multivariate location and scatter are the sample mean $\bar{\boldsymbol{x}}$ and the sample covariance matrix $\boldsymbol{S}$, i.e. the corresponding MLE estimates. These estimates are optimal if the data come from a multivariate normal distribution but are extremely sensitive to the presence of even a few outliers (atypical values, anomalous observations, gross errors) in the data. If outliers are present in the input data they will influence the estimates $\bar{\boldsymbol{x}}$ and $\boldsymbol{S}$ and further will worsen the performance of the classical multivariate procedure based on these estimates. Therefore it is important to consider robust alternatives to these estimators and actually in the last two decades much of effort was devoted to development of affine equivariant estimators which have also high breakdown point. Among the most widely used estimators of this type are the Minimum Covariance Determinant (MCD) estimator of Rousseeuw (1985), for which also a fast computing algorithm was constructed -Rousseeuw and Van Driessen (1999), the S-estimators - Davies (1987), Rocke (1996) and the Stahel-Donoho estimator introduced by Stahel (1981) and Donoho (1982) and studied by Maronna and Yohai (1995). If we give up the requirement for affine equivariance, more estimators like the one of Maronna and Zamar (2002) are available and the reward is an extreme gain in speed.

Most of these estimates became available in the popular statistical packages like $S-PLUS$, SAS, $MATLAB$ as well as in R - the packages $MASS$, $rrcov$ and recently $robustbase$. The latter intends to become the "Essential Robust Statistics" R package and strives to cover the upcoming book Maronna *et al.* (2006). Substituting the classical location and scatter estimates by their robust analogues is the most straightforward method for robustifying many multivariate procedures like principal components, discriminant and cluster analysis, canonical correlation and correspondence analysis, hypothesis testing, etc. The reliable identification of multivariate outliers (covered in an other talk) which is an important task by itself, when performed by means of robust estimators, is another approach to robustifying many classical multivariate methods.

The purpose of this talk is to review the most popular estimates of the multivariate location and scatter, to present their implementation in R as well as the accompanying graphical diagnostic tools and to compare this implementation with other statistical packages in terms of functionality and computational time. Further, the application of the robust estimates in the multivariate analysis will be illustrated on several examples - robust linear discriminant analysis, stepwise linear discriminant analysis and robust Hotelling T2 test in package $rrcov$.

References

Davies P. (1987) Asympthotic behaviour of s-estimators of multivariate location parameters and dispersion matrices, *Annals of Statistics*, 15, 1269–192.

Maronna R., Martin D. and Yohai V. (2006) *Robust Statistics*, Wiley, New York.

Maronna R. and Yohai V. (1995) The behavoiur of the stahel-donoho robust multivariate estimator, *Journal of the American Statistical Assosiation*, 90, 330–341.

Maronna R. and Zamar R. (2002) Robust estimation of location and dispersion for high-dimensional datasets, *Technometrics*, 44, 307–317.

Rocke D.M. (1996) Robustness properties of s-estimators of multivariate location and shape in high dimension, *Annals of Statistics*, 24, 1327–1345.

Rousseeuw P. (1985) Multivariate estimation with high breakdown point, in: *Mathematical Statistics and Applications Vol. B*, W.Grossmann G.Pflug I. and W.Wertz, eds., Reidel Publishing, Dordrecht, 283–297.

Rousseeuw P. and Van Driessen K. (1999) A fast algorithm for the minimum covariance determinant estimator, *Technometrics*, 41, 212–223.

Stahel W. (1981) Breakdown of covariance estimators, Research Report 31, ETH Zurich, fachgruppe fuer Statistik.