The Influence Function of Stahel-Donoho Type Methods for Robust PCA

M. Debruyne[1] and M. Hubert[1]

[1] K.U.Leuven, Department of Mathematics, W. De Croylaan 54, B-3001 Leuven, Belgium

Keywords: influence function, covariance, Stahel-Donoho, PCA, PLS

1 Stahel-Donoho

Consider a p-dimensional sample $\mathbf{X} = (x_1, \ldots, x_n)$ of size n. In this paper we will concentrate on Stahel-Donoho type estimators of covariance. By this, we mean estimators based on the Stahel-Donoho outlyingness $r(x_i, \mathbf{X})$, defined as follows (Stahel, 1981; Donoho, 1982):

$$r(x_i, \mathbf{X}) = \sup_{a \in \mathbb{R}^p} \left| \frac{a^t x_i - m(a^t \mathbf{X})}{s(a^t \mathbf{x})} \right|$$

where $m(.)$ and $s(.)$ are univariate robust estimators of location and scale. In order to obtain robust estimates of the covariance matrix, we want to concentrate on those data points with small outlyingness. We consider two options.

A first approach consists of downweighting all observations according to their outlyingness. We will call this estimator weighted Stahel-Donoho (SD_w) from now on. Several choices for the weighting function have been proposed, see Maronna and Yohai (1995), Zuo et al. (2004) and Gervini (2002).

A second approach was proposed in Hubert et al. (2005). A proportion $0 < \alpha < 1/2$ is chosen. Only the $(1 - \alpha)n$ observations with smallest outlyingness are used in the estimation. We will call this estimator Stahel-Donoho with smallest outlyingness (SD_{so}) from now on.

In this paper we derive the influence function of the SD_{so} estimator of covariance. This will allow us to consider following topics.

- Visualising the effects of outliers by plotting the influence function in the two or three dimensional case.
- Give some insight in the robustness of the estimator by calculating gross error sensitivities for general α and p.
- Calculating asymptotic efficiencies for several values of α and p.

All these results will be compared to the corresponding results for SD_w and MCD, obtained by Gervini (2002) and Croux and Haesbroeck (1999).

2 Applications in PCA and PLS

PCA is a very popular technique for analyzing multivariate data. It consists of finding orthogonal directions which maximize the variance captured in the data. These directions can be computed as the eigenvectors of an estimate of the covariance matrix. Classical PCA uses the classical sample covariance matrix to do so and thus outliers can have a very damaging effect. Recently ROBPCA, a robust PCA algorithm was proposed by Hubert et al. (2004). In this method, the SD_{so} estimator of covariance plays a crucial role.

A widely used technique for high dimensional regression is PLS. A robust version of this method was introduced by Hubert and Vanden Branden (2003). In this RSIMPLS algorithm, the SD_{so} estimator again occurs.

In our paper we will derive influence functions of both methods, using the results from Section 1. We will visualize the effects of contamination in three dimensions, giving some insight in the robustness of the algorithms. In higher dimensions, we will show that the influence functions are bounded, thus proving an important condition of the robustness of ROBPCA and RSIMPLS.

References

C. Croux and G. Haesbroeck (1999). Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis*, 71, 161–190.

D.L. Donoho (1982). *Breakdown properties of multivariate location estimators*, Qualifying paper, Harvard University, Boston.

D. Gervini (2002). Influence function of the Stahel-Donoho estimator of multivariate location and scatter. *Statistics and Probability Letters*, 60, 425–435.

M. Hubert, P.J. Rousseeuw and K. Vanden Branden (2005). ROBPCA: a new approach to robust principal component analysis. *Technometrics*, 47, 64–79.

M. Hubert and K. Vanden Branden (2003). Robust methods for Partial Least Squares Regression. *Journal of Chemometrics*, 17, 537–549.

R.A. Maronna and V.J. Yohai (1995). The behavior of the Stahel-Donoho robust multivariate estimator. *Journal of the American Statistical Association*, 90, 330–341.

W.A. Stahel (1981). *Robuste schätzungen: infinitesimale optimalität und schätzungen von Kovarianzmatrizen*, PhD thesis, ETH Zürich.

Y. Zuo, H. Cui and X. He (2004). On the Stahel-Donoho estimator and depth-weighted means of multivariate data. *The Annals of Statistics*, 32, 167–188.