Deterministic algorithms for robust covariance and regression

Tim Verdonck*
KU Leuven, Department of Mathematics, Leuven, Belgium, Tim.Verdonck@wis.KU Leuven.be

Mia Hubert
KU Leuven, Department of Mathematics, Leuven, Belgium

Peter J. Rousseeuw
KU Leuven, Department of Mathematics, Leuven, Belgium

Kaveh Vakili
KU Leuven, Department of Mathematics, Leuven, Belgium

Dina Vanpaemel
KU Leuven, Department of Mathematics, Leuven, Belgium

Many multivariate datasets contain outliers, that is data points that do not follow the pattern suggested by the majority of the data. The traditional techniques for fitting means, covariance matrices and linear regression are known to be sensitive to outliers. A single bad outlier may cause that the results are distorted so as to fit the outlier well. The same conclusion holds for other popular multivariate methods that use these estimators as building blocks (e.g. principal component analysis and discriminant analysis). Therefore various robust alternatives have already been proposed in literature.
A very robust estimator of multivariate location and scatter is the minimum covariance determinant (MCD). Computing the MCD is very hard, so in practice one resorts to approximate algorithms. Most often the FASTMCD algorithm is used. Roughly summarized, this algorithm starts by randomly drawing many random subsets and then applies so-called concentration steps (C-steps) to obtain a more accurate approximation to the MCD. The FASTMCD algorithm is affine equivariant but not permutation invariant. In this talk, we present a deterministic algorithm, denoted DetMCD, which does not use random sets.
Furthermore, we have also used the ideas from DetMCD to construct a deterministic algorithm for multivariate S-estimators of location and scatter. S-estimators for multivariate location and scatter are also highly robust and are used in several applications, such as the calculation of MM-estimators and for principal component analysis.

A very popular estimator for robust multiple regression is the least trimmed squares (LTS) estimator. Besides being highly robust, the LTS estimates are regression, scale and affine equivariant. To compute the LTS in an efficient way, the FASTLTS algorithm was developed. This algorithm is very similar to FASTMCD. It also applies a type of C-steps starting from many initial fits on random subsets. Consequently this algorithm is also not deterministic, hence not permutation invariant. Therefore, we present a deterministic algorithm for robust multiple regression, denoted as DetLTS, which does not draw random subsets.
The performances of the deterministic algorithms are compared to the corresponding "FAST algorithms" using simulated and real data sets. It is shown that the deterministic versions are permutation invariant and very close to affine equivariant. Moreover, the deterministic algorithms are in general significantly faster and sometimes more robust at highly contaminated data sets.

Key Words: Outliers, MCD, LTS, S-estimators.