Robust Estimation with FAST-MCD and FAST-LTS

Katrien Van Driessen, Peter J. Rousseeuw
UFSIA-RUCA Faculty of Applied Economics
University of Antwerp

When analyzing real data we are often faced with outliers in the data set. These outlying values can for instance be caused by measurement errors, or they may have been sampled from another population. In general we define an outlier as an observation that does not behave like the majority of the data. The classical estimators are highly influenced by such anomalous observations and so they do not yield reliable summaries of the data when outliers are present. For this reason robust methods have been developed that reduce the effect of outliers on the estimates. One of the most important challenges is to find robust estimators of the center and shape of a data cloud, since these estimators are the basis of many multivariate statistical analyses. Another very popular method is regression, which is used to explain one or more response variables by means of several regressors. The growing interest in data mining nowadays requires methods that can be computed very fast. Here we will focus on the algorithms FAST-MCD and FAST-LTS that we developed to analyze large data sets.

The minimum covariance determinant (MCD) method is a highly robust estimator of multivariate location and scatter. Its objective is to find h observations (out of n) whose covariance matrix has the lowest determinant. The MCD location estimate then is the mean of these h points, and the estimate of scatter is their covariance matrix. Because it is impossible to compute the MCD exactly if the number of observations or the number of variables is large, one generally uses approximate algorithms. However, no such fast algorithm was available, whereas nowadays large data sets are encountered more often than ever. The FAST-MCD algorithm is able to approximate the MCD for large data sets within very little time. The algorithm is based on the C-step which easily allows, starting from a currently best solution, to construct a solution with an even lower determinant. The algorithm also contains nested subsets and selective iteration as time-saving techniques. We also introduce the distance-distance plot. This plot shows robust distances (based on the robust estimates of location and scatter) versus the classical distances. This plot helps us to visualize deviating observations or unexpected structures in multivariate data.

In the regression setting it is well known that the classical least squares estimator is highly attracted by even a single outlier, and then it does not give a good fit to the linear trend of the majority of the data. Robust least trimmed squares (LTS) regression is based on the subset of h cases whose least squares fit possesses the smallest sum of squared residuals. The FAST-LTS algorithm is based on an adapted C-step which obtains a lower objective function starting from an initial trial. Nested subsets and selective iteration are included to obtain feasible computation times at large data sets. To distinguish the different kinds of outliers in the data, the diagnostic plot shows for each observation its robust residual versus its robust distance based on the explanatory variables.