

ROBUST ESTIMATORS OF EXTREME VALUES AND OUTLIERS

Shchetinin Eu.Yu., Dolgova Eu.A., Markov P.N.

Moscow State Technology University “STANKIN”,
Applied Math
Russia, 127994, Moscow, Vadkovsky per., 1, riviera-molto@mail.ru

In this thesis we propose some robust estimators and techniques for outlier detection for skewed, Pareto-type and multivariate distributions. It has been theoretically pointed out that smoothing the empirical distribution function with an appropriate kernel and bandwidth can reduce the variance and mean squared error (MSE) of some quantile-based estimators in small data sets. We apply this idea on several quantile-based estimators of location, scale and skewness and propose a robust bandwidth selection and bias reduction procedure. We use smoothed quantile-based estimators to improve the outlier detection performance of the adjusted boxplot when the data set is small.

Next, we propose an outlier detection tool for Pareto-type distributions. Classical estimators for the extreme value index, like the Hill estimator, tend to overestimate this parameter in the presence of outliers in the data. In order to measure the influence every (potentially outlying) data point has on the Hill estimator, the empirical influence function plot, which displays the influence that each data point has on the Hill estimator, is introduced. To avoid a masking effect, the empirical influence function is based on Robust Generalized Linear Model (RGLM) estimator for the extreme value index. This RGLM estimator is also used to determine high quantiles of the data generating distribution, allowing to flag data points as unusually large if they exceed this high quantile.

The final contribution of this thesis consists of the proposal of a deterministic algorithm to compute multivariate S- and MM-estimators of location and scatter. The current fast random algorithm to compute an S-estimator, which we refer to as the FastS estimator, applies improved steps to decrease the objective function in each step, similar to the FAST-MCD algorithm to compute the Minimum Covariance Determinant (MCD) estimator. Recently a deterministic algorithm for the MCD-estimator has been proposed, and we combine ideas from this method with the improvement steps from the FastS algorithm in order to construct a deterministic algorithm for multivariate S-estimators of location and scatter. We show through a simulation study that proposed S- and MM-estimators are very close to affine equivariant, and that they are permutation invariant.

References

1. *Maronna, R. A., Martin, R. D. and Yohai, V. J.*, Robust Statistics: Theory and Methods. John Wiley and Sons, 2006.