SCALABILITY PROPERTIES OF A NEW FORWARD SEARCH ALGORITHM

D. Perrotta, M. Riani, A. Cerioli, F. Torti
EC Joint Research Centre
Ispra, ITALY
and University of Parma
Parma, ITALY
e-mail: domenico.perrotta@ec.europa.eu

Abstract

We propose some computational improvements for a robust method for the identification of atypical observations known as forward search, which is based on the idea of monitoring quantities of interest, such as parameter estimates and test statistics, as the model is fitted to data subsets of increasing size. We provide a recursive implementation of the procedure which exploits the information of the previous step and we demonstrate the computational advantages of the new approach using both synthetic and real data.

1 Overview

The identification of atypical observations and the immunization of data analysis against both outliers and failures of modelling are important aspects of modern statistics. The forward search is a graphics rich approach that leads to the formal detection of outliers and to the detection of model inadequacy combined with suggestions for model enhancement. The key idea is to monitor quantities of interest, such as parameter estimates and test statistics, as the model is fitted to data subsets of increasing size.

Recent applications of the forward search include systematic outlier detection in official Census data [3], and the analysis of international trade markets [1], where important issues such as incorrect declarations, tax evasion and money laundering are at the forefront. In both these instances the number of datasets to be analyzed is of the order of hundreds of thousands, while the sample size of each dataset ranges from less than 10 observations to more than 100000. It is thus crucial to improve the computational features of the methodology and to dramatically reduce its computation time. Otherwise, on line monitoring and outlier detection, which are essential requirements for the successful implementation of statistical methods in these fields, would be unfeasible.

Therefore, the goal of this work is to provide the computational and algorithmic advances that are necessary to apply the forward search to the massive datasets arising in applications like those sketched above. In particular, we provide a recursive implementation of the forward search procedure which exploits the information of the previous step. The output is a set of efficient routines for fast updating of the model parameter estimates, which do not require any data sorting, and fast computation of

likelihood contributions, which do not require any inverse matrix or qr decomposition. It is shown that the new algorithms enable a reduction of the computation time by more than 80%. Furthermore, the running time now increases almost linearly with the sample size.

All the routines discussed are included in the FSDA toolbox for MATLAB which is freely down-loadable from Internet [2].

The scalability properties of the traditional and new forward search algorithms and of comparable robust estimators will be demonstrated on datasets of different sizes extracted from a repository of trade declarations of the customs services of the EU Member States.

References

[1] Cerioli A., Perrotta D (2013). Robust fitting of regression mixtures to contaminated trade data with high density regions. *Advances in Data Analysis and Classification*. Accepted.

[2] Riani M., Perrotta D., Torti F. (2012). FSDA: A MATLAB toolbox for robust analysis and interactive data exploration. *Chemometrics and Intelligent Laboratory Systems*, Vol. **116**, pp. 1732.

[3] Torti F., Perrotta D., Francescangeli P., Bianchi G. (2013). A robust procedure based on forward search to detect outliers in Census data. Submitted.