# Robust Analysis of Large Data Sets

Ruben Zamar (University of British Columbia) ,
Stefan Van Aelst (Ghent University)

05-06-2004–19-06-2004

Robust statistics has been developing for several decades now, and it was time for some reflection on the topic. The informal atmosphere, lengthy discussions and extended talks in our workshop permitted, amongst others things to focus on the following important questions.

- **What are the fields of application where robust statistics can make a difference?** It became clear that analysis of huge, messy data sets with many variables (the typical data-mining setting) is crucial. Applications in finance, genetics, chemometrics, etc. have been presented at the meeting.
- **Which kind of methods do we still need to develop?** From the discussions and talks, it became clear that analysis of datasets that contain several types of data (numerical, categorical, ordinal, etc.), robust statistical inference beyond point estimation, and cleaning of data matrices deserve more attention.
- **Do we need to put into questions some of the foundations of robustness?** In the robustness community, high breakdown point and equivariance properties have always been at a central place, but it is not always clear that these properties are required in typical data mining applications. Moreover it is not obvious how to define important robustness notions such as breakdown point in non-standard settings.

From the discussions at the workshop it became clear that the interplay between robustness and data mining will be an important direction of future research with many applications. In data mining the focus is on extracting valuable information from large databases. Revolutionary progress in digital data acquisition and storage in recent years has resulted in the creation of huge databases. Supermarket transactions, credit card usage, telephone calls details, internet traffic, corporate statistics, astronomical data, gene expression data, medical and clinical data are all examples of such databases. In fact, the production and accumulation of digital databases is occurring at a faster rate than our ability to comprehend and use them. One possible reaction to this avalanche of digital data would be to dismiss them as electronic junk. However, many people including the organizers and most participants of this workshop believe that these databases contain valuable knowledge which can be mined (found, extracted and used).

In the process of mining the data - as in a true mining operation - we go through the following typical main steps:

| | | |
|---|---|---|
| Step I | **Defining mining goal:** | defining the particular type of structure (or structures) we wish to find. |
| Step II | **Scoring mining results:** | deciding how to quantify - score - the success of a given structure in realizing the mining goal. |
| Step III | **Getting it done:** | designing and implementing an algorithm to optimize the scoring scheme from Step II. |

Since Statistics and Data Mining are both concerned with the analysis and modelling of data and methods to perform these tasks, there is a big overlap between these two disciplines. There is a clear parallel between data mining Steps I, II and III and Statistical Model Building, Model Fitting and Computing. The main difference lies in the fact that, given the magnitude of the datasets encountered in data mining applications, Step III has to be highly automated and run with non or little human intervention. Data miners have always used statistical tools and statistician are now showing an interest in Data Mining problems. The interactions between the two disciplines will be very beneficial to both of them.

It has been discussed by the participants of our workshop (and generally agreed) that there is an opportunity for the application of robust methods and ideas in Data Mining. It is possible, for example that some useful patterns apply to the majority (but not the totality) of the data. Such patterns may never be found by classical statistical methods that attempt to fit the complete dataset. Robust algorithms, on the other hand, search for suitable subsets of the data and therefore can find these partial patterns. But classical robust procedures are computationally intensive and do not scale well to large datasets normally encountered in data mining applications.

Several examples shown at the workshop clearly illustrated that there is a need for robust data analysis techniques that can handle large data sets. However, many of the robust methods that are available cannot be applied directly to large data sets due to practical and theoretical reasons. Practically, robust methods are computationally so demanding that running them on large data sets is not feasible in a reasonable amount of time. From the theoretical viewpoint many robust methods are not suitable for large, high dimensional data sets because they are based on the concept of outlying objects (rows in the data matrix). Most available robust methods treat the measurements for all variables of one object as the basic processing unit. Each object is classified as “good” or “outlying” and if an object is considered outlying, then all measurements for that object are downweighted together. This approach works well for low dimensional data sets because it leads to equivariance properties that are often considered desirable. In high dimensions, on the other hand, it is not reasonable anymore to consider all measurements of an object as deviating from the majority if some of them are outliers. Indeed, often only 1 or a few of the measurements are contaminated while all other measurements of the object are not. In fact, if every variable has a small probability of producing a contaminated measurement, then the probability of having a completely clean object decreases rapidly as the dimension increases. In high dimensions, we can thus be confronted with data sets that contain only few completely clean objects. This violates the basic assumption underlying most available robust procedures: good points form the majority of the data. For high dimensional data sets it is therefore more natural to consider **outlying cells** instead of **outlying objects**. A cell is the measurement for one object and one variable. Robust statistics can make relevant contributions to the field of data mining by developing methods and techniques based on outlying cells that are better suited to handle the problems encountered when analyzing large data sets. Research in this direction should focus on developing methods and algorithms that are maybe less refined from the statistical viewpoint but are extremely fast to compute and scale well with growing sample size and dimension.