Statistical Analysis of Complex data sets with Robust Statistical methods

http://www.firstlight.cn 2007-04-20

11. April 2007，Robust statistical analysis methods capable of dealing with large complex data sets are required more than ever before in almost all branches of science. The European Science Foundation's three-year SACD network, which was completed in December 2006, developed new methods for extracting key structural features within the data. Such features can include outlying values that may be particularly significant within the increasingly large and complex data sets generated in financial markets, medical diagnostics, environmental surveys, and other sources.

"Outliers often indicate the most interesting data points, like polluted areas for environmental data, or irregularities in online monitoring of patients," said SACD chair Christophe Croux. On this front the programme has almost completely achieved its objectives, according to Croux. "A lot of work has been done in developing new methods, especially for analyzing large data sets, that can cope with outlying atypical values. This resulted in a number of publications related to the subject of the network".

Particular progress has been made detecting outliers in multivariate time series, Croux added. This is a significant development for a number of analysis and monitoring applications involving measurements of different but related quantities that vary over time. Among many such applications are: monitoring of telecommunication networks to assess how performance and reliability are affected by events such as upgrades, surges in demand, and local link failures; monitoring noise in the vicinity of an airport; modeling the behaviour of financial markets in response to geopolitical events; and tracking the condition of patients in intensive care via several measurements such as pulse rate, blood pressure, lung water etc.

Without robust analysis methods it is easy to miss significant outlyers in such multivariate data. In some cases the outlyers only show up clearly when considering all the variables together, and yet may indicate something significant that could easily be missed, such as a sudden deterioration in a critical patient's condition.

SACD has also advanced the field of chemometrics, which is the application of multivariate analysis methods to data of chemical interest, with some of the developments now implemented in software written by members of the network. The same principles have been applied to analysis of risks of stock investments, and measuring volatility of financial markets.

In some cases it is desirable to eliminate outlyers from data sets in order to identify the most likely response of a particular variable to different events. Within SACD, a method was developed to do this for analysis of the relationship between various economic parameters and the yield of stocks. For this it is necessary to concentrate on the bulk of the data rather than the exceptions or outlyers. "In order to do so we have to identify these extreme observations in order to downweight or reject them from the computations," said Croux. When there are multiple variables this is more difficult, and one of the major achievements of SACD has been to find new ways of condensing and summarizing the data in such a way that the main structure of the data can be retrieved, making it also easier to detect the outliers.

Croux admits there is more work to be done, particularly in dealing with highly complex data sets, and with problems involving many variables and small sample sizes. "Important steps to be taken include robust methods that can deal with categorical data and missing values."

SACD has laid the ground for progress in all these areas, having stimulated interest within its three workshops by presenting data from real research in progress, rather than artificial samples. "We got much more interest than expected from colleagues in other fields. This was due to the interesting network workshops, where cutting edge scientific research was presented," said Croux.

Most important of all, presented material in the workshops held in 2004, 2005 and 2006 was often work in progress, leading to exchange of ideas and the initiation of joint research projects among the partners. In this way the work of SACD will continue and expand after the Network itself is over.

The European Science Foundation is an association of 75 member organisations devoted to scientific research in 30 European countries. Since its inception in 1974, it has coordinated a wide range of pan-European scientific initiatives.

存档文本