Mathematisches Forschungsinstitut Oberwolfach

Report No. 39/2009

DOI: 10.4171/OWR/2009/39

Challenges in Statistical Theory: Complex Data Structures and Algorithmic Optimization

Organised by
Rudolf J. Beran (Davis, CA)
Claudia Klüppelberg (TU München)
Wolfgang Polonik (Davis, CA)

August 23rd – August 29th, 2009

Abstract. Technological developments have created a constant incoming stream of complex new data structures that need analysis. Modern statistics therefore means mathematically sophisticated new statistical theory that generates or supports innovative data-analytic methodologies for complex data structures. Inherent in many of these methodologies are challenging numerical optimization methods. The proposed workshop intends to bring together experts from mathematical statistics as well as statisticians involved in serious modern applications and computing. The primary goal of this meeting was to advance the mathematical and methodological underpinnings of modern statistics for complex data. Particular focus was given to the advancement of theory and methods under non-stationarity and complex dependence structures including (multivariate) financial time series, scientific data analysis in neurosciences and bio-physics, estimation under shape constraints, and high-dimensional discrimination/classification.

Mathematics Subject Classification (2000): primary: 62xx, *secondary:* 60Bxx, 60Gxx, 60Hxx, 60Jxx, 90Cxx, 92xx.

Introduction by the Organisers

The workshop *Challenges in statistical theory: Complex data structures and algorithmic optimization*, organised by Rudolf Beran (Davis, CA), Claudia Klüppelberg (TU München) and Wolfgang Polonik (Davis, CA) was held August 23rd – August 29th, 2009. This meeting was well attended by 49 participants with diverse geographic, demographic and disciplinary representation.

The theme of the conference addressed the challenges to modern Statistics created by the ongoing emergence of novel, large, and complex data types. Human ability to collect data through sophisticated electronic technologies has outstripped human ability to distill information from the data. Resolving the situation requires, among other things, fundamental new developments in statistical theory and algorithms. Traditional probabilistic studies of statistical procedures remain an important tool but no longer suffice. Data is arguably not random in the sense of probability theory; data may reside naturally on a manifold or other algebraic structure; the procedure under study may be very complicated; and probability modeling in some modern problems, such as classification of highly structured data, has not been effective. Emerging new types of data include: single molecule observations; complex simultaneous recording of several neurons; the outcomes of computer experiments; high dimensional observations of brain waves that need to be processed in real time (if possible); or high-frequency financial data.

To date, the most powerful statistical methodologies have been developed for data that resides in a Euclidean space. Emerging data types pose a variety of questions that include: On what algebraic structure does the data naturally reside? On this algebraic structure, can we develop statistical methodologies that address the questions posed by those who collected the data? In particular, can we devise for such data analogs of successful statistical methodologies for Euclidean data? Meeting such challenges requires communication among those most involved with the new types of data, those with the expertise to identify suitable mathematical formulations, those who have thought deeply about abstract statistical inference, and those who seek to devise new paradigms for statistical reasoning beyond probability modeling of the data.

Thus, a fundamental task for Statistics is to develop powerful theoretical tools that engender and validate effective methodologies for the analysis of modern data types arising in a variety of fields. To do so first requires gaining sufficient disciplinary and mathematical insight into the new underlying data structures. The workshop brought together experts from mathematical statistics and the statistical sciences. The primary goal was to address the foregoing challenges by broadening the mathematical underpinnings of modern statistics. The secondary goal was to foster cross-fertilization between the core of statistics and the statistical sciences.

Workshop participants presented a range of novel data structures and of methodologies proposed for their analysis. Mathematical advances were exhibited, for instance in the area of model selection for functions in high dimensional spaces, in estimation under heavy tails, or in estimation under shape restrictions. Tools, such as Malliavin calculus, for the large sample analysis of certain statistics were discussed. Methodological advances in the areas of modeling of nonstationarity, the construction of confidence intervals for classification error, or the testing of functional autoregression were treated. Algorithmic and/or computational issues are inherent in many of these challenges, and many of the presentations addressed this aspect.

Summary: Complex high dimensional data is rather the norm than the exception in modern statistics, and modeling or analyzing such complex data is a huge challenge. In order to properly understand these approaches effective mathematical techniques are necessary. Making advanced methodology feasible in practical applications usually also requires devising sophisticated optimization/algorithmic methodologies. New paradigms beyond probability modeling are needed to validate complicated statistical procedures. Balancing all of these ingredients is a fundamental challenge. Substantial progress in these directions requires input from various sides. The workshop brought out the issues and made significant contributions to the program outlined.