Mathematisches Forschungsinstitut Oberwolfach

Report No. 13/2011

DOI: 10.4171/OWR/2011/13

Mini-Workshop: Level Sets and Depth Contours in High Dimensional Data

Organised by
Mia Hubert, Katholieke Universiteit Leuven
Jun Li, University of California, Riverside
Wolfgang Polonik, University of California, Davis
Robert Serfling, University of Texas, Dallas

February 27th – March 5th, 2011

Abstract. Extraction of information about the distribution underlying a high-dimensional data set is a formidable, complex problem dominating modern nonparametric statistics. Two general strategies are (i) to extract merely qualitative information, such as modality or other shape information, and (ii) to consider relatively simple inference problems, such as binary classification. One approach toward (i) and (ii) is based on measuring qualitative information via mass concentration functions. Another approach is based on multivariate depth functions and inherently addresses issues of robustness. Having different orientations and aims, these approaches have evolved in parallel with little interaction. Yet they both in common implicitly involve level set estimation as a major tool. This mini-workshop was the first serious attempt to study and exploit such interconnections between these approaches. Researchers from both areas exchanged ideas toward forging a novel, synergistic approach that fruitfully strengthens the roles of mass concentration and depth methods in statistical inference for multivariate data. Foundations for level set estimation as a general statistical method were explored. Deeper understanding of the so-called generalized quantiles approach was pursued. Application to binary classification, a pervasive problem in modern statistics, received intensive special attention.

Mathematics Subject Classification (2000): Primary: Nonparametric Inference 62G99, Secondary: Multivariate Analysis 62J99.

Introduction by the Organisers

The mini-workshop *Level Sets and Depth Contours in High Dimensional Data*, organized by Mia Hubert (Katholieke Universiteit Leuven), Jun Li (UC Riverside), Wolfgang Polonik (UC Davis) and Robert Serfling (UT Dallas) on February 27–March 5th, 2011 brought together 18 participants with diverse geographic, demographic and with research expertise in level sets and/or depth methodology.

Statistical methodology for high-dimensional data problems faces many challenges, many of them relate to *geometry*. One general strategy for dealing with related dilemmas is to consider less complex goals. Here one approach is to obtain *qualitative* information about *shape*, with monotonicity, modality, or *mass concentration* of the underlying distribution being specific instances. Another approach involves the use of *multivariate quantiles* and *depth functionals.* In fact, the notion of a quantile is closely related to the notion of *outlyingness*, which in turn connects with *robustness* of multivariate statistical methodology.

Interestingly, the two above-mentioned approaches both entail *the estimation of level sets*: (a) level sets of depth functionals, or depth contours, provide a measure outlyingness of outlyingness of multivariate data; (b) several mass concentration functions of a multivariate distribution can be considered as functionals of level sets of the corresponding probability density function; (c) information about modality of a distribution is reflected in the shape of level sets of the probability density function; (d) in a classification context, the Bayes decision boundary of the optimal classifier is given by a level set of the regression function.

A second general strategy in dealing with the challenges posed by high dimensionality is to confine to relatively simple statistical inference procedures. A very important example is the *classification or discrimination problem.* Again we find level set estimation as an underlying tool: for example, the optimal (Bayes) classifier is a level set of the corresponding regression function (conditional expectation). In fact, typically (binary) classifiers are characterized by a decision boundary which may be represented as $\{x : g(x) = 0\}$ for some decision function g. In other words, if an observation falls into the level set $\{x : g(x) \geq 0\}$ then it is classified into one class, and otherwise it is classified into the other class.

Although different in their orientation and goals, and having evolved with relatively little interaction, depth-based methods and the mass concentration approach are connected via technical commonalities revolving around the general theme of *level set estimation* and through certain applications such as the classification problem. Formal investigation and systematic exploitation of such connections among these quite distinct statistical settings would be novel and fruitful. It is the chief target of the proposed workshop, and other spin-offs are anticipated as well.

This mini-workshop brought together representatives of the depth and mass concentration groups along with interested statisticians from related areas. This was a first serious attempt to forge a new synergy yielding a deeper understanding of level set and depth contour estimation and their applications and to spawn new research directions.