Mathematisches Forschungsinstitut Oberwolfach

Report No. 49/2006

Qualitative Assumptions and Regularization in High-Dimensional Statistics

Organised by
Lutz Dümbgen (Bern)
Jon A. Wellner (Seattle)

November 5th – November 11th, 2006

Abstract. Important and exciting developments are currently underway in nonparametric statistics involving inter-play between qualitative constraints, penalization, and regularization methods. Some of these developments are taking place on the theoretical side (with connections in the direction of empirical process theory), while other parts of the development are occurring on the algorithmic and approximation theory sides. This workshop brought together researchers from several of these groups to exchange ideas and problems, to probe further research directions.

Mathematics Subject Classification (2000): primary: 62xx, secondary: 41Axx, 52Axx, 65Kxx, 90Cxx.

Introduction by the Organisers

This workshop was well-attended with 47 participants from Europe and overseas, among them many promising young scientists. While most participants are working in mathematical statistics, several participants are experts in approximation theory or fields of application such as astrophysics or econometrics, too. The participants exchanged ideas, discussed new developments and established new projects and interactions for the subsequent tasks.

Traditional nonparametric statistics and new trends. Nonparametric statistics has undergone dramatic changes during the last two decades. At first the focus shifted from permutation and rank testing in classical settings such as comparison of two univariate samples to multi- and even infinite-dimensional problems such as density estimation and regression. Here new results and techniques from empirical process theory, a very active research area in itself, played a prominent role.

At present statisticians working on nonparametrics are facing three kinds of problems, among others: On the one hand, the research focused strongly on point estimation, whereas in applications people are in need of tests and confidence sets. A second problem is the curse of dimensionality. Roughly speaking, the number of unknown parameters for reasonable approximating models grows exponentially with the dimension. This problem results in rather slow rates of convergence in higher dimensions. In addition, many estimation problems are inverse in the sense of involving indirect measurements and being ill-posed.

Qualitative Assumptions. For all three problems introducing qualitative assumptions is turning out to be a successful strategy with further potential. That means, in many situations, restrictions on the underlying function parameters such as e.g. monotonicity, concavity/convexity, or upper bounds on the number of local extrema may be used to enhance the performance of point estimators substantially and to replace quantitative smoothness assumptions which are difficult to justify. In addition, imposing such constraints enables the construction of nonparametric tests and confidence sets, sometimes even without relying on asymptotic expansions.

Computation and Regularization. In order to deal with the qualitative assumptions algorithmically, techniques for constrained optimization come into play. Sometimes it turns out that standard solutions from optimization theory such as, for instance, quadratic programming, are not efficient for statistical purposes, and alternative procedures such as the pool-adjacent-violators algorithm have been developed by statisticians. Naturally, regularization methods are used in this context, too. Regularization methods themselves are a well-known tool for treating inverse problems. In statistics they are also important in order to produce "sparse" estimators, i.e. estimators which are easier to interpret because of few non-zero parameters, few local extrema or other characteristics. In fact, in many fields of application the underlying parameter itself is assumed to be sparse, at least approximately. This is in fact the intrinsic reason why nonparametric curve estimation is possible at all.

Dimensional Asymptotics. More recently some authors showed how to use regularization successfully in regression problems with sparse parameters but of dimension p growing almost exponentially with the number n of observations. Considerations of this type are increasingly important, showing new trade-offs between flexibility of models and stability of estimation. One obvious example is the analysis of gene expression data, where the number of parameters (gene fragments) is in the range of a few hundred to several thousand, while sample sizes are rarely larger than a few hundred. Here approaches such as penalized logistic regression turn out to be very promising.

Informal Sessions. In addition to the regular talks (see the abstracts below), we organized two informal evening sessions with the following talks:
Arnold Janssen: Regions of alternatives with high and low power for goodness-of-fit tests,
Angelika Rohde: Adaptive goodness-of-fit tests based on signed ranks,

Melanie Birke: Estimating a convex function in nonparametric regression,
Kaspar Rufibach: The log–concave density estimator as a smoother,
Nicolai Bissantz: Nonparametric testing in noisy inverse problems.