Research statement by Georgy Shevlyakov

Directions of research

My interests are mainly concerned with **the theory and applications of robust statistical methods** and with **the mathematical methods and technologies of data analysis**. The researches aim both at theoretical aspects and real-life needs of data and signal processing. Robust statistics and data analysis are relatively new disciplines dealing with statistical inferences under the conditions of uncertainty of probability distribution models. I introduced several new concepts, published one book and about fifty papers on these topics. The monograph "Robustness in Data Analysis: criteria and methods" written in collaboration with Nikita Vilchevski summarizes our former results obtained in the eighties-nineties. Recently, I began to study the problems of robust signal processing, namely, robust detection of signals and robust fusion, as well as some problems in queueing theory and Kalman filtering. However, now I am mostly interested in developing of robust statistical methods and focusing at the areas of robust correlation analysis and robust multivariate statistics with their applications to data and signal processing.

Robust statistics

Robustness of the least absolute values *(LAV)* method I began to deal with the problems in robust statistics from the very beginning of my academic career when being a student. In my M.Sc. degree thesis, the dependence of optimal linear combinations of order statistics on the shape of an underlying distribution was studied and robust versions of the classical Grubbs and Dixon tests for detection of outliers in the data were proposed. In my Ph. D. dissertation, entitled as "Robust Estimation and Detection of Signals Based on the *LAV* Method", the global robust properties of the *LAV* method (a robust alternative to the classical least squares (*LS*) method) were studied: basing on the classical results from the approximation theory in the L_1-norm metric, I obtained the explicit boundaries upon the breakdown points of the *LAV* estimators of regression, namely, for the algebraic and trigonometric polynomial type covariates. In this work, I collaborated with Eugene Gilbo and Igor Chelpanov.

Huber's minimax approach in robust statistics In general, the minimax principle aims at the worst situation for which it suggests the best solution. Huber's minimax approach presumes the following two stages: (i) a least favorable (informative) distribution minimizing Fisher information is obtained; (ii) **t**he maximum likelihood estimator for this least favorable distribution is used. For the problems of robust estimation of location, I obtained the least favorable distributions in the several new distribution classes with bounded variances and subranges (such restrictions are typical in practice) with the subsequent optimal maximum likelihood estimators. In particular, a new parametric family of distributions called the Weber-Hermite distributions was introduced as the solution to the Euler-Lagrange equation in the variational problem of minimizing Fisher information. This family comprises the normal and the Laplace distributions as the limit cases, and the corresponding optimal minimax estimator of location has the following three branches: (i) with relatively small distribution variances or with relatively short distribution tails (in the sense precisely defined), it is the sample mean or the *LS* estimator; with relatively large variances or with relatively heavy distribution tails, it is the sample median or the *LAV* estimator; (iii) and with relatively moderate variances, it is a compromise between the *LS* estimator and the *LAV* estimator which can be approximately described as the *Lp*-norm estimator ($1<p<2$). This solution is robust and close to the well-known Huber's solution for the class of heavy-tailed ε-contaminated normal distributions due to the presence of the *LAV* estimator branch and more efficient than Huber's for the short-tailed distributions due to the presence of the *LS* estimator branch. In other words, the additional information on the relative weight of distribution tails may significantly improve the quality of estimation. These results were extended on the solutions of regression problems and they also partially work within autoregressive models. I collaborated with Nikita Vilchevski on these topics.

Stability of the least favorable distributions In general, the solutions of variational problems strongly depend on the regularity conditions imposed upon the class of sought optimal functions. In Huber's minimax approach, the determination of a least favorable distribution is the crucial step, and a natural question arises whether a solution to the variational problem of minimizing Fisher information is stable (robust) to the violations of regularity conditions. Considering several classes of lattice distributions as the discrete analogs of the corresponding classes of continuous distributions widely used in robust statistics, e.g., Huber's class of heavy-tailed ε-contaminated normal distributions, I showed that the least favorable lattice distributions fully preserve the qualitative structure of their continuous prototypes. It is important that these results do not depend on which of the possible two lattice analogs of Fisher information is used in the set-up: the first analog is directly derived from the classical functional by extracting its main part while the limit transition from a continuous to a lattice distribution; to derive the second analog, it is necessary to obtain a new form of the classical Cramér-Rao inequality for lattice distributions and, following the classical recipe, to take the denominator of the upper bound upon the variance of an unbiased estimator as the required lattice analog of Fisher information. In this work, I also collaborated with Nikita Vilchevski.

Robust correlation I began to deal with the problem of robust estimation of correlation in the beginning of the eighties after the pioneer works by S. Devlin, R. Gnanadesikan and J. Kettenring in which they considered the classes of robust estimators of a correlation coefficient based on the direct robust counterparts of the sample correlation coefficient and on the robust estimators of the principal component variances. I introduced the following two new classes of robust estimators of a correlation coefficient, namely, the class of estimators based on robust linear regression and the class of two-stage estimators based on the preliminary rejection of outliers in the data with the subsequent application of the sample correlation coefficient to the rest of the data. In order to reveal most prospective estimators of correlation, I analyzed the asymptotical and finite samples performance of the typical representatives (totally about 20 estimators) of all the aforementioned classes of robust estimators of correlation. It turned out that the best robust estimators belong to the class of estimators based on the robust estimators of the principal component variances: namely, *the trimmed and the median correlation coefficients.* This experimental observation was confirmed by the following precise result: I showed that those estimators are the minimax variance (in the Huber sense) estimators of a correlation coefficient for ε-contaminated bivariate normal distributions. Thus, *the trimmed and the median correlation coefficients* are the correlation analogs of such well-known robust estimators of location as the trimmed mean and the sample median, respectively. These robust estimators of a correlation coefficient were used for the element-wise robust estimation of correlation matrices, as well as for robust estimation of their eigenvalues and eigenvectors. In these work, I collaborated with Vladimir Pasman, Tatiana Khvatova, Jae Won Lee, and Nikita Vilchevski.

Data analysis

A probability-free approach to the choice of an optimization criterion In mathematical statistics, the choice of an optimization criterion is predetermined by the choice of a probability distribution model. In the conditions when the observed data samples are unique or small and, therefore, the basic requirement of applicability of the mathematical methods of statistics, i.e., stability of frequencies (see H. Cramér "Mathematical Methods of Statistics", Princeton, 1946) cannot be verified, the choice of an optimization criterion sometimes can be justified by the other arguments. In our book (G. Shevlyakov, N. Vilchevski "Robustness in Data Analysis: criteria and methods", VSP, 2002), we consider the set-up connected with replacing of a data collection by a unique characteristic quantity. To a certain extent, this characteristic called the "typical" representative (a measure of location in the Cauchy sense) is equivalent to the initial data and thus it can be interpreted as an estimator for all of the data. It is defined as the solution of the problem of minimization of some measure of total discrepancy between observations and their "typical" representative. It is shown that such a measure of discrepancy must satisfy certain *a priori* postulated natural requirements towards the properties of a "typical" representative. These requirements mainly follow from metrological restrictions. The latter commonly are the requirements of equivariancy to the translation, scale, orthogonal, and monotonic transformations of the data. We show that taking into account such metrological requirements we can narrow the admissible classes of measures of discrepancy and in some cases to reduce them to parametric dependencies. In particular, the requirement of scale equivariancy results in the *Lp*-norm estimates with arbitrary values of *p,* the requirement of affine equivariancy leads to the *LS* method, and the requirement of monotonic equivariancy yields the *LAV* method.

Data compression via smoothing the sample quantile function The quantile function, the inverse to the distribution function is a convenient tool in exploratory data analysis, in some aspects even more convenient than the distribution function itself. The Bernstein polynomials are widely used in the theoretical studies on convergence processes of approximations to functions continuous on [0, 1], since these polynomials have a rather simple structure and provide uniform convergence in the Weierstrass theorem sense. In order to obtain a smooth estimate of the sample quantile function, I suggested to use the Bernstein polynomials and some their generalizations. The asymptotic properties of these estimates were studied. A two-stage procedure of data compression based on the preliminary smoothing of the sample quantile function by the Bernstein polynomials with the subsequent application of the Chebyshev-Padé approximations of these smoothed estimates was proposed. It exhibited a rather good performance reducing the dimensionality from $n=150$ down to the 5 coefficients of the Chebyshev-Padé approximation of the Bernstein polynomial estimate for the sample quantile function. Here I collaborated with Nikita Vilchevski.

A low-complexity bivariate boxplot Basing on the highly robust estimators of location, scale and correlation, namely, the sample median, the median absolute deviation and the median correlation coefficient, I have proposed a low-complexity bivariate boxplot for compression and visualization of the bivariate data. This construction naturally generalizes the well-known univariate technique proposed by John Tukey. Currently, Peter Filzmoser from Technical University of Vienna (Austria) and I are working on the comparative analysis of the performance of that and related methods.

Signal processing

Robust detection of signals The aforementioned results on robust estimation in models with bounded noise variances obtained within Huber's minimax approach have been effectively used in robust detection of signals. The designed robust detectors are simultaneously highly resistant to the presence of outliers in the data and highly efficient in approximately Gaussian noise. This area of studies still remains very attractive for me. Currently, I study application of the so-called redescending *M*-estimators (with score functions tending to zero with increasing values of an argument) to robust detection. Since redescending *M*-estimators of location outperform conventional bounded *M*-estimators of Huber's type, it is very likely that so it will be in detection. In this work, I collaborate with Stephan Morgenthaler from Ecole Polytechnique Fédérale de Lausanne (Switzerland), as well as with Kiseon Kim and Jin Tae Park from Gwangju Institute of Science and Technology (Korea).

Kalman filter estimates Kalman filters are widely used in signal processing for estimation of state parameters of static and dynamic systems. The practical needs of mechatronics signal processing require the methods and algorithms providing optimal and robust fusion of local estimates in multisensor filtering problems. In these work concerned with some extension of classical results and application of robust methods and algorithms to Kalman filtering, I collaborate with Vladimir Shin, Kiseon Kim and Nguen-Na Viet from Gwangju Institute of Science and Technology (Korea).

Networks queueing algorithms Now this area of research is especially important for communications signal processing. I was involved in these works before my current affiliation, and the experience of work with queueing models, namely with priority queueing and generalized processor sharing with applications to ATM networks and differentiated services in the Internet, helped me much to adapt to the required directions of research in communications signal processing. I collaborated with Konstantin Avrachenkov and Nikita Vilchevski on this topic.

Selected applications

The automatic system for diagnostics of cardiac diseases The participation (as a member of a big team of medical doctors, engineers, mathematicians) in creation of the mathematical and statistical software for the automatic system of the cardiac disease diagnostics was my most significant work in real-life applications. I joined that team in 1981 and have been working on the related problems for more than 10 years. During this work, numerous problems of medical data processing, e.g., statistical estimation and classification, mathematical modeling of the professional experience of medical doctors, were treated and

solved. The experience of collaboration with the specialists from different areas gained during this work gave me the understanding of the technology and of the spirit a team work. The first version of the system was created in 1986, later it has been modified and implemented in medical practice providing the quality of the cardiac diagnostics at the level of 95% of correct diagnoses.

Sudden cardiac death risk factors Statistical analysis of the sudden death risk rate, detection of the sudden cardiac death risk factors and study of their dependence on the meteorological and solar activity (in collaboration with Prof. Lev Chireikin from St. Petersburg Institute of Cardiology), was the most cha1lenging and intriguing problem in applied statistics I ever faced. The exposure of sudden cardiac death (SCD) risk factors (RF) is a rather old problem. There are many researches treating this problem and similar questions mainly studying the influence of cardio-pathological factors (blood pressure, atherosclerosis), psychological factors (stresses), social factors (smoking, alcoholism), etc. Here we considered only the RFs connected with meteorological and solar factors. This problem is important since: (i) the mortality rate by coronary heart diseases is about 50% of the general mortality; (ii) the rate of SCDs is about 70% of the coronary heart mortality. The initial data was unique: the daily measured number of SCDs (all clinically verified), meteorological factors (the average temperature, its daily increment, the daily increment of pressure, wind speed) and solar factors (the terrestrial magnetic activity, Wolf number, area of sunspots, integral solar activity index, number of sun flares) in Arkhangelsk (one of the northern Russian cities, about 400,000 of population) for 1983-1985, totally 1096 days. By applying the methods of robust statistics (robust correlation and factor analysis), it was found out that the most dangerous (murderous) are the solar activity factors with the instantaneous impact, especially the integral solar activity index, but not the terrestrial magnetic activity and meteorological factors whose impacts are delayed in time and secondary in importance. The further analysis showed that the expected values of murderous solar factors with the instantaneous impact are almost unpredictable (what is rather natural), and that, in general, only the passive protection of very weak and easily vulnerable patients in hospitals is possible.

Future research plans

These plans are mainly concerned with robust statistics and data analysis technologies including the extension of robust methods on the problems of multivariate statistical and time series analysis, normalizing data transformations, as well as some problems in general statistics with their applications to data and signal processing in electrical engineering:

Robust multivariate statistics: robust correlation, robust principal components analysis, and detection of multivariate outliers in the data.

Robust time-series analysis: robust singular spectrum decomposition (SSD).

Data analysis: bivariate boxplots techniques, methods for normalizing the data.

General statistics: estimation in finite distribution models.

Applications: signal processing (robust estimation and detection of signals) in electrical engineering, robust statistical classification methods in pattern recognition.

A final remark on my research philosophy

I consider statistics as the intermediate territory between mathematics and “real life” sciences (natural and social) with the problems mostly originating from the latter and the methods taken from the former.

By the force of things and as a matter of taste, I definitely prefer clearly (both in the “real life” science and in the mathematical senses) posed statistical problems which can be tackled and solved using classical and powerful mathematical methods of analysis, optimization, calculus of variations, and special functions.