Raport Badawczy

Research Report

RB/70/2006

Book review:
**Statistical data analysis
based on the L1-norm
and related methods
by Yadolah Dodge (Editor)**

P. Nowak

Book review:

STATISTICAL DATA ANALYSIS BASED ON THE L1-NORM AND RELATED METHODS

by

Yadolah Dodge (Editor)

The volume is a selection of papers, presented to the fourth International Conference on Statistical Analysis Based on the L1-Norm and Related Methods, held in Neuchâtel, Switzerland in 2002. The theory and applications of statistical methods based on L1-norm are a promising fields of statistics. The evidence of this fact is a wide variety of papers, which describe many real-life applications of presented methods. The volume contains 36 papers grouped in seven parts.

The first two papers of Part one, "All About Quantiles", concern quantiles, M-quantiles and the spread of multivariate distributions. The notions considered by authors are illustrated by a series of examples. The next paper discusses a "spatial scale curve", which is a two-dimensional characterization of the spread of multivariate distributions. The curve is a plot of a "volume" functional based on "central regions" of increasing size. In the fourth paper of this part the author presents some results on order statistics of locally dependent processes and their applications to quantiles of linear processes, linear combinations of order statistics, and a multivariate extension. Some possibilities of asymptotic theory of sample quantiles are considered in the next paper. Examples with detailed calculations are good illustrations of the presented theory. The last two papers of Part one describe interesting applications of the quantile methods to the selection indices for university admissions and exploring transition data related to unemployment duration.

Part two, "Financial Statistics and Time Series", is a collection of six articles. The papers show the importance of applications of statistical methods in finance. The first paper "How to be Pessimistic: Choquet Risk and Portfolio Optimization" is a review of new approaches to the theory of risk assessment. It is shown that pessimistic portfolio optimization may be formulated in terms of quantile regression. The author of the next paper considers Expected Shortfall as an alternative to the currently most popular measure, called value-at-risk. Expected Shortfall is characterized as the smallest coherent and law invariant risk measure to dominate value-at-risk. An application of binary

quantile regression to the problem of credit scoring is the subject of the third paper. The quantile regression discrimination is compared favorably to probit and heteroskedastic probit discrimination. The subsequent paper is devoted to prediction of 0-1 events for short- and long-memory time series. The introduced method is illustrated by an example of its application to daily values to stock market indices. The next paper of Part two is dedicated to problems concerning smoothing parameter choice in quantile regression, used in statistical analysis of stock returns. Some possibilities of bandwidth selection, based on cross validation, generalized cross validation, plug in and double smoothing are presented. A similar problem is considered in the last paper of this part. It is addressed to the bandwidth selection for kernel estimates of the marginal probability distribution of a time series that is a transformation of a stationary zero mean Gaussian process. In the paper a bandwidth selection procedure is proposed.

Part three, "Estimation, Testing and Characterization", contains ten papers. The first one is devoted to the least orthogonal absolute deviations (LOAD) estimates, closely related to estimation of errors-in-variables models, construction of robust alternatives to principal components, and detection of rank deficient observations in regression. The second paper contains considerations on L1-derivatives and the score functions which are important tools in statistical inference, the maximum likelihood estimation, score tests and locally or asymptotically optimal rank tests. Authors of the third paper find a family of M-estimates of regression, which minimize the asymptotic variance at the central model, subject to a bound on the maximum bias over contamination neighborhoods, and they presents some numerical results. The next paper is dedicated to robust bootstrap for S-estimators of multivariate regression. In particular, it is shown that confidence intervals for regression coefficients based on the robust bootstrap have good performance when compared to other approaches. The subsequent article contains theoretical results on M-tests for detection of structural changes in regression with proofs and simulations. To obtain approximations for critical values, modified permutational arguments are used. Similar studies are presented in the next article, where the hypothesis that the observations have the same distributions versus the alternative that the distribution of observations changes at some point is verified. The paper: "A Class of Probability Metrics and its Statistical Applications" is very interesting from theoretical and practical point of view. The space of probability measures with metrics, belonging to a considered class, is isometric to a subset of a Hilbert space. Moreover, corresponding minimum-distance estimators are studied. The next paper is a discussion of the on line fitting procedures, related to maximum likelihood estimation, for Gaussian and Laplacian distributions. The consecutive article is dedicated to the estimation of covariance matrices based on multivariate sign and rank vectors. Finally, in the last article of Part three, for the multivariate (elliptically symmetric) one-sample location problem, a class of optimal procedures is developed with application of "Tyler angles".

The next part of the book, "Deep in the Data", deals with contributions

related to data depth. The first paper is a description of methodologies based on L1-depth in regression and outlier detection. In particular, two methods for linear regression are studied. The second paper is devoted to a description of the variation of halfspace regions under perturbation, as well as behavior of the perturbed halfspace median in the case of absolutely continuous centrosymmetric distributions. In the third article of this part multivariate trimmed means, based on data depth, are introduced and their properties are studied.

Classification problems are considered in Part five. The first paper is dedicated to p-distances in connection with non-hierarchical models in classification. The next article contains a comparison of the classification method based on the support vector machine with the one based on the regression depth method. The third paper presents a robust clustering method, which is based on modified Weiszfeld algorithm for the multivariate median and data depth. The method is applied to gene expression data. The next article discusses the concept of median extension for metric structures, especially for spaces with semi-distance of L1-type. The last paper of this part is devoted to an application of several text classification models to the mining of aviation inspection reports. It shows that text classification can be a useful approach to aviation safety decision support systems.

The subsequent part of the book, entitled "Density Estimation and Image Processing" contains three articles. The first paper compares estimators based on L1 Markov random field with wavelet-based estimators. Moreover, an interesting application to financial time series is presented. The next paper has an empirical character and it is addressed to nonparametric regression based on regularization. It compares the elastic and plastic splines based on total variation penalties. The third paper develops a compression scheme for microarray images, using a bitplane coding strategy. The development is based on an extension of the JPEG2000 lossless standard and a robust L1 vector quantizer.

Part seven, "Regression Models in Environmental Studies", consist of two papers, describing environmental applications. The first one examines two regression models for overdispersed count data. The models are applied to a study of the association between daily emergency hospital visits for respiratory diseases and the levels of air pollution. The second paper is devoted to an analysis of complex regression models, which is being applied to the pollution and mortality problems in São Paolo.

The papers presented in volume are the evidence that L1-norm methods play an important role in many basic fields of modern statistics

The volume provides an interesting exposition of the theory and applications of statistical analysis based on the L1-norm. This part of statistics is very important for applications, and, simultaneously, it is very valuable for researchers, since the presented theory can be developed in many different, promising directions.

Piotr Nowak

Yadolah Dodge, ed., *Statistical Data Analysis Based on the L_1-norm and Related Methods.* Series: SIT – Statistics for Industry and Technology. Birkhäuser Verlag, Basel – Boston – Berlin, XII+454 pages, 2002. ISBN 978-3-7643-6920-0. Price (hardcover, special offer until January 2007): 60.94 .